DOCUMENT RESUME

ED 109 722 CS 501 092

TITLE	Status Report on Speech Research: A Report on the Status and Progress of Studies on the Nature of Speech, Instrumentation for Its Investigation, and Practical Applications, January 1 - March 31, 1975.
INSTITUTION	Haskins Labs., New Haven, Conn.
REPORT NO	SR-41-(1975)
PUB DATE	Mar 75
NOTE	236p.
EDRS PRICE	MF-$0.76 HC-$12.05 PLUS POSTAGE
DESCRIPTORS	*Cognitive Processes; *Communication (Thought Transfer); Educational Research; Higher Education; Language Development; Listening Skills; *Research Methodology; *Speech; Speech Skills; Stuttering; *Theories; Vision
IDENTIFIERS	*Status Reports

ABSTRACT

This report is one of a regular series on the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. "Preliminaries to a Theory of Action with Reference to Vision" attempts to describe how the contents of vision may relate to the process of action. "On the Relationship of Speech to Language" reviews several theories on the relationship between language and verbal communication and language and the mental functions. "Pitch in the Perception of Voicing States in Thai: Diachronic Implications" examines changes in stop consonant voicing in the Thai family of languages by seeking new information on acoustic cues in modern Thai. "A Combined Cinefluorographic-Electromyographic Study of the Tongue During the Production of /s/: Preliminary Observations" explores the interrelationships of muscle activity, tongue movement, and the resultant acoustic signal. And "The Stuttering Larynx: An EMG, Fiberoptic Study of Laryngeal Activity Accompanying the Moment of Stuttering" investigates the hypothesis that the most common cause of stuttering is the glottis. (RB)

Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes every effort to obtain the best copy available. nevertheless, items of marginal reproducibility are often encountered and this affects the quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDRS). EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made from the original.

ED109722

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRO-
DUCED EXACTLY AS RECEIVED FROM
THE PERSON OR ORGANIZATION ORIGIN-
ATING IT. POINTS OF VIEW OR OPINIONS
STATED DO NOT NECESSARILY REPRE-
SENT OFFICIAL NATIONAL INSTITUTE OF
EDUCATION POSITION OR POLICY

SR-41 (1975)

Status Report on

SPEECH RESEARCH

A Report on
the Status and Progress of Studies on
the Nature of Speech, Instrumentation
for its Investigation, and Practical
Applications

1 January - 31 March 1975

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports.)

CS 501 092

ACKNOWLEDGMENTS

The research reported here was made possible in part by support from the following sources:

National Institute of Dental Research
Grant DE-01774

National Institute of Child Health and Human Development
Grant HD-01994

Research and Development Division of the Prosthetic and Sensory Aids Service, Veterans Administration
Contract V101(134)P-71

Office of Naval Research, Information Systems Branch
Contract N00014-67-A-0129-0001

Advanced Research Projects Agency, Information Processing Technology Office, under contract with the Office of Naval Research, Information Systems Branch
Contract N00014-67-A-0129-0002

United States Army Electronics Command, Department of Defense
Contract DAAB03-75-C-0419(L 433)

National Institute of Child Health and Human Development
Contract NIH-71-2420

National Institutes of Health
General Research Support Grant RR-5596

HASKINS LABORATORIES

Personnel in Speech Research

Alvin M. Liberman,* President and Research Director
Franklin S. Cooper, Associate Research Director
Patrick W. Nye, Associate Research Director
Raymond C. Huey, Treasurer
Alice Dadourian, Secretary

Investigators

Arthur S. Abramson*
Fredericka Bell-Berti*
Gloria J. Borden*
Earl Butterfield[1]
James E. Cutting*
Christopher J. Darwin[2]
Ruth S. Day*
Michael F. Dorman*
Peter Eimas[3]
Jane H. Gaitenby
Thomas J. Gay*
Katherine S. Harris*
Philip Lieberman*
Leigh Lisker*
Ignatius G. Mattingly*
Paul Mermelstein
Seiji Niimi[4]
Lawrence J. Raphael*
Donald P. Shankweiler*
George N. Sholes
Michael Studdert-Kennedy*
Michael T. Turvey*
Tatsujiro Ushijima[4]

Technical and Support Staff

Eric L. Andreasson
Dorie Baker*
Elizabeth P. Clark
Cecilia C. Dewey
Janneane F. Gent
Donald S. Hailey
Harriet G. Kass*
Diane Kewley-Port*
Sabina D. Koroluk
Christina R. LaColla
Roderick M. McGuire
Agnes McKeon
Terry F. Montlick
Susan C. Polgar*
Loretta J. Reiss
William P. Scully
Richard S. Sharkany
Edward R. Wiley
David Zeichner

Students*

Mark J. Blechner
Susan Brady
John Collins
David Dechovitz
Susan Lea Donald
G. Campbell Ellison
Donna Erickson
F. William Fischer
Carol A. Fowler
Frances J. Freeman
Gary M. Kuhn
Andrea G. Levitt
Roland Mandler
Robert F. Port
Robert Remez
Philip E. Rubin
Helen Simon
James M. Vigorito

---

*Part-time
[1] Visiting from the University of Kansas, Lawrence.
[2] Visiting from the University of Sussex, Brighton, England.
[3] Visiting from Brown University, Providence, R. I.
[4] Visiting from University of Tokyo, Japan.

CONTENTS

I. Manuscripts and Extended Reports

Preliminaries to a Theory of Action with Reference to Vision -- M. T. Turvey . . . 1

Two Questions in Dichotic Listening -- Michael Studdert-Kennedy . . . 51

On the Relationship of Speech to Language -- James E. Cutting and James F. Kavanagh . . . 59

Rise time in Nonlinguistic Sounds and Models of Speech Perception -- James E. Cutting, Burton S. Rosner, and Christopher F. Foard . . . 71

Phonetic Coding of Words in a Taxonomic Classification Task -- G. Campbell Ellison . . . 95

On the Front Cavity Resonance, and Its Possible Role in Speech Perception -- G. M. Kuhn . . . 105

Synthetic Speech Comprehension: A Comparison of Listener Performances with and Preferences Among Different Speech Forms -- Patrick W. Nye, Frances Ingemann, and Lea Donald . . . 117

Testing Synthesis-by-Rule with the OVEBORD Program -- Frances Ingemann . . . 127

Stress and the Elastic Syllable: An Acoustic Method for Delineating Lexical Stress Patterns in Connected Speech -- Jane H. Gaitenby . . . 137

Is it VOT or a First-Formant Transition Detector? -- Leigh Lisker . . . 153

Pitch in the Perception of Voicing States in Thai: Diachronic Implications -- Arthur S. Abramson . . . 165

Facial Muscle Activity in the Production of Swedish Vowels: An Electromyographic Study -- Katherine S. Harris, Hajime Hirose, and Kirstin Hadding . . . 175

A Combined Cinefluorographic-Electromyographic Study of the Tongue During the Production of /s/: Preliminary Observations -- Gloria Borden and Thomas Gay . . . 197

Velar Movement and Its Motor Command -- T. Ushijima and H. Hirose . . . 207

The Stuttering Larynx: An EMG, Fiberoptic Study of Laryngeal Activity Accompanying the Moment of Stuttering -- Frances Freeman and Tatsujiro Ushijima . . . 217

II. Publications and Reports . . . 231

III. Appendix: DDC and ERIC numbers (SR-21/22 - SR-39/40) . . . 235

I. MANUSCRIPTS AND EXTENDED REPORTS

Preliminaries to a Theory of Action With Reference to Vision*

M. T. Turvey[+]
Haskins Laboratories, New Haven, Conn.

Of the distinction his own efforts had done much to foster, Magendie commented in 1824:

> The organs which concur in muscular contraction are the brain, the nerves, and the muscles. We have no means of distinguishing in the brain those parts which are employed exclusively in sensibility, and in intelligence, from those that are employed alone in muscular contraction. The separation of the nerves into nerves of feeling and nerves of motion is of no use: this distinction is quite arbitrary (cited in Evarts, Bizzi, Burke, Delong, and Thach, 1971:111-112).

More recently this point of view has been expressed in a different but closely cognate fashion by Trevarthen (1968:391): "Visual perception and the plans for voluntary action are so intimately bound together that they may be considered products of one cerebral function."

In the light of such remarks, it is curious that theories of perception are rarely, if ever, constructed with reference to action. And while theories of perception abound, theories of action are conspicuous by their absence. But it must necessarily be the case that, like warp and woof, perception and action are interwoven, and we are likely to lose perspective if we attend to one and neglect the other; for it is in the manner of their union that the properties of each are rationalized. After all, there would be no point in perceiving if one could not act, and one could hardly act if one could not perceive.

Of course, history has not been remiss in comments on the relation between perceiving and acting. From the time of Aristotle it has been taught that the motor system is the chattel of the sensory system. Nourished by the senses, the

---

*To be published in Perception, Action, and Comprehension Towards an Ecological Psychology, ed. by R. Shaw and J. Bransford. (Pontiac, Md.: Erlbaum, in press).

[+] Also University of Connecticut, Storrs.

Acknowledgment: The preparation of this paper was supported by a Guggenheim Fellowship awarded to the author for the period 1973-1974. I thank N. S. Sutherland for kindly providing facilities and assistance at the Laboratory of Experimental Psychology, University of Sussex, Brighton, England. Some of the ideas expressed in this paper were presented at the Conference on Perception, Action, and Comprehension, held at the Center for Human Learning, University of Minnesota, Minneapolis, in August 1973.

[HASKINS LABORATORIES: Status Report on Speech Research SR-41 (1975)]

motor system obediently expresses in automatonlike and relatively uninteresting fashion the cleverly contrived ideas of the higher mental processes, themselves offshoots of the sensory mechanisms. In this view, action is interpretive of the sensory mind and thus, in principle, problems of coordinated activity are secondary to and (if we assume an associative link between sensory and motor) independent of problems of perception. It has also been taught, usually with less fervor, that perception is a disposition to act: to perceive an event is to be disposed to respond in a certain way. Modification of this view leads to a constructive theory of mind in which it is argued that higher mental processes in addition to perception are skilled acts that reflect the operating principles of the motor system. In short, experience is constructed in a fashion intimately related to the construction of coordinated patterns of movement. So far as action assumes primary importance in this approach to mind, we would expect its proponents to put great store by the analysis of coordinated motions. However, where motor-theoretic interpretations have been forwarded to account for perception and the like, statements of how acts are actually produced have been either absent or trivial (e.g., Sperry, 1952; Bartlett, 1964; Festinger, Burnham, Ono, and Bamber, 1967; Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). Curiously, action-based theories of perception and of mind in general have been advanced on a nonexistent theory of action.

Thus, it seems that the theory of action deserves more attention than it has received and that the interlacing of the processes of perceiving and acting is a problem we can perhaps no longer afford to ignore. This essay is a preliminary and speculative response to these reproofs. Its purpose is twofold: first, it seeks to identify a set of basic principles to characterize the style of the action system in the production of coordinated activity; and second, it attempts, in a rough and approximate way, to describe how the contents of vision may relate to the processes of action. To a significant degree, the ideas expressed in this paper derive, on the one hand, from the work of Bernstein (1967) and the Russian investigators who have followed his intuitions, and, on the other hand, from the analysis and amplification of the Russian views by Greene (1971a, 1971b). We begin our inquiry by illustrating an equivalence between problems of action and the more heralded problems of perception and cognition (cf. Turvey, 1974).

THE CONSTANCY FUNCTION IN ACTION, AND ACTION AS CONSEQUENCE

A visually presented letter A can occur in various sizes and orientations and in a staggering variety of individual scripts. Yet in the face of all this variation, the identification of the letter remains, for all intents and purposes, unaffected.

This phenomenon of constancy is not limited to the domain of perception, but is equally characteristic of action. Thus, the letter A may be written without moving any muscles or joints other than those of the fingers. Or, it may be written through large movements of the whole arm with the muscles of the fingers serving only to grasp the writing instrument. Or, more radically, one can write the character without involving the muscles and joints of either arms or fingers, by clenching the writing instrument between one's teeth or toes. It is evident that a required result can be attained by an indefinitely large class of movement patterns.

On examination of the phenomenon of constancy in action we might raise the query: How can these indefinitely large classes of possible movement patterns be stored in memory? The answer is that they are not. Clearly, I do not have on record in memory all possible temporal sequences of all possible configurations of muscle motions that write A; indeed, I have yet to perform them and by all accounts I never shall. The essential question about our A-writing task, therefore, can be stated more fundamentally: How can I produce the indefinitely various instantiations of A without previous experience of them?

In response to this question, let us turn our attention to linguistic theory. A departure point for transformational grammar is that our competency in language is such that we can produce and understand a virtually infinite number of sentences. As Weimer (1973) has pointed out, there are echoes of Plato's paradoxes in Chomsky's (1965) claim that our competence in language vastly outstrips our experience with it. Chomsky's claim is motivated by the observation that experience with a limited sample of the set of linguistic utterances yields an understanding of any sentence that meets the grammatical form of the language. To explain this competency is, for Chomsky (1966), a central problem in the Theory of Language. But given the points advanced above, the constancy function in action is likewise indicative of a competency that exceeds prior learning. The child, we may note, learns to write A under conditions that restrict her to a small subset of the very large set of A-writing movements. But she is able subsequently to write A with practically any movement pattern she chooses, i.e., she can write A in novel ways. A-writing is creative in the sense that language is creative.

The search for a workable account of the creativity manifest in language has led transformational grammarians to what can be aptly described as "the explanatory primacy of abstract entities" (Hayek, 1969). The idea is that the speaker-listener has at his disposal an abstract system of rules or principles, referred to as the deep structure, that allows him to generate and to understand an indefinitely large set of sentences, referred to as the surface structure. This distinction, drawn in linguistic theory, between deep and surface structure will prove relevant to our analysis of action in two important respects. The first is the idea that deep structure is far removed from surface structure; grammarians argue that although the deep structure determines the surface structure, it is not manifested in the surface structure. The second is that the child must come to determine the nature of the underlying deep structure from a limited experience with surface structures. Chomsky and his colleagues assume that the child essentially "looks through" the utterances she hears to the abstract form behind those utterances. The child is said, therefore, to construct a theory of the regularities of her linguistic experience. Similarly, our hypothetical child learning to write the letter A must determine from her limited experience with the set of A-writing movements a theory of how to write A. Thus, we may conclude that the ability to write A in indefinitely various ways is based on procedures that are abstract and generative, like the grammar Chomsky has in mind for language. Others have sought similar parallels between action and grammar (e.g., Lenneberg, 1967).

There is an interesting upshot to this discussion of action constancy. We generally say that an abstract representation, a concept, underlies our ability to recognize indefinitely various A's. Let us call this the perception concept of A. Now clearly we may propose that there is an action concept of A underlying our ability to write A in indefinitely various ways. So are there in general two different kinds of structures, two different classes of concepts--one specific to perceptual events, the other specific to action events? In short, is the constancy function in perception achieved in ways fundamentally different from the constancy function in action? If it is, then the construction of theories of how we identify events (see Neisser, 1967)--theories of the perception concept--can proceed virtually independent of the construction of theories of the action concept. On the other hand, if the constancy function is treated in the same way in both perception and action, that is, if there is only one class of appropriate structures or only one class of appropriate procedures for achieving constancy, then the theory of identification and the theory of production ought not to be considered separately. In this view, which I suspect is the more viable, any account of constancy in perception must also be an account of constancy in production--a perceptual account of constancy must be potentially translatable into an action account of constancy. If such a translation is in principle implausible, then we may suppose that the account is incorrect.

The reader's attention is drawn in this preamble to one other important aspect of action--its relation to "consequence." An act modulates environmental events, but philosophers have found that they cannot conceptually distinguish between occurrences that are actions and occurrences that are consequences (see Care and Landesman, 1968). A typical argument from language usage might go like this: George kicks the football (of the round kind) and scores the goal that wins the championship. Now we could say that George kicked the football and that a consequence of his action was that a goal was scored. Or we could say, just as appropriately, that George scored a goal with championship-winning consequences. "Scored the goal," therefore, can be viewed either as consequence or as action. We may wish for criteria to determine which occurrences should receive an action label, and which occurrences should receive a consequence label. Unfortunately, the criteria that have been advanced have not met with any degree of universal approval.

The failure to distinguish conceptually between action and consequence is understandable from the viewpoint of Bernstein (1967). He comments:

> Whatever forms of motor activity of higher organisms we consider... analysis suggests no other guiding constant than the form and sense of the motor problem and the dominance of the required result of its solution, which determines, from step to step, now the fixation and now the reconstruction of the course of the program as well as the realization of the sensory correction (p. 133).

The implication is that an action plan as a statement of consequences is not a static structure but a structure that is, by virtue of processes we shall discuss below, continually becoming. Yet in all of its phases of change, phases that constitute a tailoring of the plan to the current kinematic and environmental contingencies, the essential character of the action plan remains invariant. What is to be achieved, what is to be the consequence of the evolving pattern of motions, persists from the conception of an act through its evolution to its completion.

The arbitrariness of distinguishing between action and consequence parallels the arbitrariness of distinguishing between perception and memory. As William James (1890) observed, and as others concur (e.g., Gibson, 1966a), the traveling moment of present time is not a razor's edge and no one can identify when perception ends and memory begins. The distinction between action and consequence is as much a will-o'-the-wisp as the distinction between perception and memory.

THE DOMAIN OF ACTION CONCEPTS

For present purposes, we shall entrust ourselves to the view of concepts as functions (Cassirer, 1957). Thus we may represent an action concept such as that for A-writing as A(x) and explore the nature of the variable x that enters into this function. We perform this exercise in order to identify some fundamental characteristics of the action system. Let us assume that the elements entering into A(x) are a proper subset of the set of elements that enter into any rule for coordinated activity. And, in addition, that coordinated activity is under the management of an "executive system" and that the character of the elements entering into A(x) and any other action function are mirrored in the character of (or constraints on) this system.

One view of the executive is that expressed in the traditional piano or push-button metaphor. In this metaphor, muscles are represented cortically in keyboard fashion, one muscle per key, and central impulses to the muscles are held to be unequivocally related to movement. The essence of the view is that the executive instructs each muscle individually. At the outset we may question the worth of this metaphor simply on the ubiquity of reciprocal innervation: the intricate and extensive interrelation among muscles makes it both arduous and wasteful to instruct them singly. But more importantly, we can argue (as did Bernstein, 1967) that there cannot be an invariant relation between innervational impulses and the movements they evoke.

Consider the movement of a single limb segment in relation to a fixed partner and under the influence of a single muscle. The differential equation describing this situation is of the form:

$$I\frac{d^2\alpha}{dt^2} = f\left(E, \alpha, \frac{d\alpha}{dt}\right) + g(\alpha)$$

where I is the inertia of the limb segment, α is the angle of articulation, E the innervational level of the muscle, and f and g are the functions determining, respectively, the muscle force and gravitational force acting on the limb segment.

If we take $E = E(\alpha, d\alpha/dt)$, that is to say, independent of time and simply a function of position and velocity, then the equation reduces to that for a movement of a limb indifferent to central influences; in brief, an instance of central paralysis. If, for contrast, we assume that the excitation of a muscle is solely a function of a centrally predetermined sequence and independent of the peripheral variables of position and velocity, that is, $E = E(t)$, then the equation is that of a system insensitive to, or ignorant of, changes in local conditions. Obviously, it is more judicious to argue that $E = E(t, \alpha, d\alpha/dt)$, in which case the fundamental equation can be written:

$$I\frac{d^2\alpha}{dt^2} = f\left[E\left(t, \alpha, \frac{d\alpha}{dt}\right), \alpha, \frac{d\alpha}{dt}\right] + g(\alpha)$$

Solutions to equations of this kind depend on the initial conditions of integration. The implication, therefore, is that in order to obtain the same movement for various values of α and $d\alpha/dt$, different innervational states E will be needed. In a word, the relationship existing between impulses to the muscle and the movement of the single limb segment is equivocal: same impulses may produce different movements and different impulses the same movement.

We continue Bernstein's argument by noting that in the temporal course of moving a limb segment changes occur in the force of gravity [which is related by a function $g(\alpha)$ to the angle of articulation] and in other external forces operating on the limb and that these changes affect E. Now suppose that the limb segment traces out a rhythmical motion. This rhythmical motion can be identified with a function relating the required forces at the joint to time. However, another function can be identified relating forces at a joint to time, and the forces in this case correspond to the changes in the external force field. As a result, the sequence of impulses to the muscle can be interpreted as determining a mapping of the function generated by the variations in the external field over time to the desired function. Now suppose that the same rhythmical motion is traced out with the hand holding on separate occasions (a) a hammer, (b) a baton, and (c) a can of beer. The function relating the changes in the external force field to time will differ in each instance even though the pattern of the rhythmical movement is unchanged. In each of the three instances a different mapping would be required from the function generated by the external force field to the desired function specifying the rhythmical pattern. The import of this, as Bernstein (1967:20-21) points out, is that the sequence of impulses to the muscle "cannot maintain even a remote correspondence" to the factual form of the movement.

A third criticism of the push-button metaphor is that if the executive behaved in the fashion suggested, instructing each muscle individually, then it would be called upon to manage the enormous number of degrees of freedom that the motor apparatus attains "...both in respect to the kinematics of the multiple linkages of its freely jointed kinematic chains, and to the elasticity due to the resilience of their connections--the muscles. Because of this there is no direct relationship between the degree of activity of muscles, their tensions, their lengths, or the speed of change in length" (Bernstein, 1967:125). Herein lies a fundamental principle that simply states that the number of degrees of freedom of the system controlling action is much less than the number of mechanical degrees of freedom of the controlled system (Kots, Krinskiy, Naydin, and Shik, 1971). A homely example illustrates the point: try writing a letter, e.g., W, while simultaneously making circular motions with a foot. An experimental illustration is provided by Gunkel (1962): when one makes movements of different rhythms simultaneously with the two hands, the amplitude of the movements performed by one of the hands is modulated by the frequency of the movements performed by the other. Thus, it is not difficult to demonstrate that the number of degrees of freedom of the executive is very small; on the push-button metaphor it would have to be very large. We can conclude, therefore, on three counts, that the executive does not, or indeed cannot, control individually each motor unit or even each muscle participating in a complex act.

One consequence of the conclusion that the executive system does not control muscles singly is that it need not be apprised of peripheral details, since such information would be irrelevant. In this light, let us take another look at the equation for the movement of a single limb segment. In that equation the innervational impulse is expressed as a function of time, angle of articulation (muscle length), and velocity, i.e., $E = E(t, \alpha, d\alpha/dt)$. But if the executive is stripped of the responsibility for instructing individual muscles and if it is ignorant of the current, precise details of the external force field, then clearly executive instructions are not written in the form relevant to that field, i.e., in the form $E(t, \alpha, d\alpha/dt)$.

Moving a single limb segment rhythmically requires an action plan and we may suppose that executive instructions spell out that plan (in the sense of defining the contours and timing of the movement) through a sequence of impulses of the form $E(t)$. The action plan and impulses of the form $E(t)$ must correspond, or so it would seem, to the factual form of the movement, in contrast to impulses of the form $E(t, \alpha, d\alpha/dt)$, which on the above account bear no such relationship to the movement. Thus we see that the action plan (the deep structure) is dissimilar to the innervational signals issued to the muscles (the surface structure) and these signals in turn are dissimilar to the movement that evolves: "...it is as if an order sent by the higher center is coded before its transmission to the periphery so that it is completely unrecognizable and is then again automatically deciphered" (Bernstein, 1967:41). In general, if impulses of the form $E(t)$ are close to the action plan and hence close to the actual form of the movement, then those impulses of the form $E(t, \alpha, d\alpha/dt)$ are close to the muscles and to the actual forces operating at the joint complexes. On this view, the mapping of $E(t)$ to $E(t, \alpha, d\alpha/dt)$ identifies the evolution of an act; in particular, it identifies the adaptation of an action plan to the prevailing field of external forces.

But if the executive does not control individual muscles, then what does it control? In response to this question, students of action (e.g., Bernstein, 1967; Gelfand, Gurfinkel, Tsetlin, and Shik, 1971) propose that the executive charge is to control the modes of interaction of lower centers. These, it is argued, are capable, through the systems that they govern, of producing a coordinated movement pattern in a relatively autonomous fashion.

Consider a commonplace, coordinated activity such as running. There are lower centers that control individual limbs, with each center asserting particular relations among the components of the limb that it controls. Thus, the interaction between these centers determines the coordinated motion of the limbs, and the problem of coordination in running becomes for the executive a problem of intercenter coordination (Shik and Orlovskii, 1965). Let us pursue this example in more detail because it is representative of a mode of organization that we will entertain as characteristic of the action system.

We have evidence that mechanisms inherent in the segmental apparatus of the mammalian spinal cord can initiate and maintain flexion-extension or stepping movements of the limbs in the absence of afferent participation (Eldred, 1960). Apparently these segmental pattern generators determine the fundamental form of flexion-extension activity, but they do not specify in detail the actual spatial and temporal characteristics of the motion (Engberg and Lundberg, 1969). It is the role of afferent information, enumerated through autonomous (reflex) structures (and of tuning influences from above, as we shall see later), to supply the requisite spatial and temporal details and thus to tailor the basic pattern to the field of external forces. A small leap now takes us to the assertion that walking and running can be attributed to a relatively simple executive instruction that sets into characteristic motion the entire segmental apparatus and which, in itself, is deficient in information about the actual strategic order of necessary muscle contractions (cf. Evarts et al., 1971).

This mode of organizing action achieves the following. First, it resolves the degrees of freedom problem noted above by apportioning relatively few degrees to the executive level but relatively many to the subsystems whose activities the executive regulates. (Since it is the subsystems that must deal with the vagaries of kinematic linkages and muscles.) Second, and related, it reduces the detail required of the executive instructions, for with autonomous lower centers those instructions do not have to be coded for the individual muscle contractions that will ultimately occur.

In overview, what has emerged is the understanding that the element entering into the design of an act is typically not an individual muscle but a group of muscles functioning cooperatively together. We have good reason to speculate that the reflexes may well comprise the "basis" of the set of all such functional groupings and hence of the infinitely large set of all acts (Easton, 1972a). A "basis" is a mathematical structure found in the theory of vector spaces. It is defined as a linearly independent (nonredundant) set of vectors that under the operations of addition and scalar multiplication spans the vector space. Essentially, a "basis" contains the minimum number of elements that are required to generate all members of the set.

We have several reasons for identifying the set of reflexes as the "basis" for action. First, reflex systems are not independent entities that function in isolation. On the contrary, there are a multiplicity of functional relations among reflexes and other structures. Second, virtually every reflex observed experimentally and clinically is an instance of a reasonably complex configuration of motions often elicitable by a single stimulation. Third, reflex systems are under very effective and often complex control by supraspinal structures (cf. Eccles and Lundberg, 1959; Kuno and Perl, 1960; Evarts et al., 1971). And fourth, reflexes are obviously purposeful and adaptive, and they may be organized and modulated flexibly by means of the operations of ordering, summing, fragmentation, and through their "local sign" properties (Easton, 1972a). Collectively, these characteristics of reflexes suggest that "...the neuronal mechanisms which have been studied as reflex arcs can be utilized in a variety of ways by virtue of the interaction between reflex pathways and by the action of control systems that are present, even at the level of the spinal cord segment. The dichotomy between reflex control and central-patterning control of movement may in this sense be artificial"(Evarts et al., 1971:62).

Through the provision of reflexes, evolution has supplied a partial answer to the degrees of freedom problem. We might now suppose that a further reduction in the burden of control is achieved ontogenetically through the gathering together of reflexes into larger functional units (cf. Paillard, 1960; Pal'tsev, 1967b; Gelfand et al., 1971). We shall refer to reflexes and functional combinations of reflexes as "coordinative structures" [a term borrowed from Easton (1972a) but used here with greater latitude].[1] Of cardinal importance to this essay is the assumption that a closely knit functional combination of reflexes performs as a relatively autonomous unit; by this assumption, relative autonomy is a fundamental property of coordinative structures, whether large or small.[2]

---

[1] One motivation for bringing reflexes and functional combinations of reflexes under the single heading "coordinative structures" is the assumption that for the activation of either a single reflex or a single functional combination of

In sum, we have seen that the executive does not construct acts from individual muscle contractions. What we now infer is that acts are synthesized from a set of coordinative structures for which the reflexes constitute a basis.

We return now to the question of the variable entering into action concepts of the form A(x). The executive does not deal in muscles, so muscle properties (length, tension) can be ruled out. The executive does deal in coordinative structures (at least so we may argue), but these similarly cannot be the elements we seek. An action concept such as that supporting A-writing is indifferent to functional groupings of muscles in the same way that it is indifferent to to individual muscles. However, the analogy drawn above between the set of reflexes and a "basis" in vector space theory provides a clue to the answer. To reiterate: a basis is a subset of a set of elements which, when acted upon by suitable operations, generates the entire set of elements. We assume, therefore, a repertoire of operations that modify and relate the coordinative structures so as to produce any and all acts. Thus we may conjecture that the elements entering into an action concept are the operations defined over the set of coordinative structures. In this sense an action concept is analogous to a mathematical operator, a function whose domain is a set of functions, of which differentiation is a classical example.

---

reflexes, one degree of freedom of the control system is enough (see Kots et al., 1971). In regard to functional combinations, it is important to recognize that new tasks may often require the discovery of new combinations and their establishment as single functional units. In very large part acquiring a skill is, as Bernstein (1967) would have expressed it, a problem of reducing the degrees of freedom in the action structures being regulated. The elegant and instructive experiments of Kots and Syrovegnin (1966) addressed the question of how the action system manages the large degrees of freedom manifest by a system of multiple links. The participant's task in these experiments was to flex or extend his wrist and to flex or extend his elbow. The investigators observed that, in the main, the joints moved in coupled fashion and that the two rates of change of joint angles maintained one of three to seven constant ratios. These constant ratios were not determined by the mechanical link of the joints; rather they appeared to be determined by a system of control in the form of a functional link between motor centers innervating the flexors and extensors of the joints.

[2] Consider insect flight. The evidence suggests that it is not due to a built-in structural system of simple segmental reflex loops nor to any flight center yet identified. Rather, it seems that there is a functional system of distributed oscillators--autonomous pattern generators--which on receipt of the appropriate nonphasic input are coupled together as a unit which then operates autonomously in a preset fashion (Weiss-Fogh, 1964). Walking may use some of the very same oscillatory structures as flying, but for locomotion on the ground they would be mutually coupled in a different way (cf. Wilson, 1962) to form a different autonomous unit.

THE ORGANIZATION OF THE ACTION SYSTEM

The foregoing account identifies two particularly important properties of the action system. First, acts are produced by fitting together structures each of which deals relatively autonomously with a limited aspect of the problem. Second, the action plan is stated crudely "in three-dimensional kinematic language" (Gelfand et al., 1971), yet the actual pattern of motions is precise in displacement, speed, and time of occurrence. To achieve this measured performance, the differentiation of an action plan must proceed through multiple stages of computation in which needed details emerge gradually. Patently, a computation of details over time is inelegant and inefficient for a system that has a limited repertoire of skills, but it is preferred for a system called upon to solve novel action problems posed by ever-varying kinematic and environmental conditions.

We commonly classify a system that behaves in this fashion as hierarchic, a classification that is certainly suggested by the unqualified use of the term "executive" in the preceding discussion. By a hierarchy we understand that an executive at the highest level of a decision tree makes the important decisions and spells out the fundamental goals. Decisions on the details are left to the immediately subordinate structures, which in turn leave decisions that they cannot make, for whatever reason, to even lower structures. This general strategy is repeated until the final remaining decisions are made by the lowest structures in the decision tree.

The crucial property of a substructure in a hierarchy is that in the perspective of a higher level it is a dependent part, but in the perspective of a lower level it is an autonomous whole. Koestler's (1969) term "holon" expresses this whole-part personality of hierarchic substructures; a holon is defined as "a system of relations which is represented on the next higher level as a unit, that is, a relatum" (Koestler, 1969:200). We may question, however, the notion explicit in the concept of a hierarchy that the direction of the whole-part personality of substructures is immutable. Certainly from the viewpoint of the "geometry" of anatomical arrangements certain structures may appear as dependent parts of other structures, and a compelling argument may well be made for the immutability of this relation in the peripheral reaches of the neural mechanisms supporting action and perception. Yet, from a computational viewpoint in which we emphasize "knowledge" structures rather than anatomical structures, the relation between any two structures need not be fixed; either may treat the other as a relatum, or subprocedure, depending on the problem to be solved at a given moment. This commutability of "subordinate" and "executive" roles, of "lower" and "higher," is expressed in the related interpretations of biological systems as "coalitions" (von Foerster, 1960; Shaw, 1971; Reaves, 1973) or "heterarchies" (Minsky and Papert, 1972). In these interpretations, management of the action system would not be the prerogative of any one structure; many structures would function cooperatively in the framing of action plans and desired consequences, although not all structures need participate in all decisions (Reaves, 1973). Furthermore, while it is certainly the case that the action system has very definite and nonarbitrary (anatomical/computational) structures, in these interpretations the partitioning of these structures into agents and instruments and the specification of relations among them is arbitrary. Any inventory of basic constituent elements and relations is equivocal (Reaves, 1973). Decentralization of control and arbitrariness of partitionings are not alien notions to students of action theory (e.g., Bernstein, 1967; Greene, 1971b) as is evident from Greene's apologia:

> The "executive" and "the low-level systems" will occur frequently.... These terms are simply abbreviations for what I really mean: any two subsystems, one of which, at the moment, in respect to the task under consideration, is behaving like an executive relative to the other. The systems are not unique, and their relation is not immutable: a "lower" part of the nervous system might, for instance, at some time behave like an executive relative to some "higher" part (Greene, 1971b:2-3).

ACTION AS HETERARCHIC

Perception and action contrast in that the tasks of the former are to digest, abstract, and generalize, while the tasks of the latter are to spell, concretize, and particularize (Koestler, 1969). One is the mirror image of the other. For the sake of argument and to facilitate comparisons with perception, let us say that the "input" to the action system is an intention (e.g., to pick up a cup, to write one's name). (We respectfully ignore the problem of how an intention is determined and in addition we give due recognition to the likelihood that some of the structures responsible for determining an intention may also be responsible for its translation into an action plan and for the plan's subsequent differentiation.) Therefore, an intention is an "event" for the action system in the way that, say, a scene is an event for the visual, perceptual system.

Taking a leaf from artificial intelligence research on visual perception, we may say that action involves knowledge domains or abstract representations--where a representation is defined as a set of entities, a description of the relations among them, and a description of their attributes (Minsky and Papert, 1972; Sutherland, 1973). Thus, for the perception of scenes portrayed in two dimensions, we may identify, as examples, (1) a Lines Domain in which "bars, picture-edge, vertex, end, mid-point" are the entities; "join, intersect, collinear, parallel" are the relations; and "brightness, length, width, orientation" are the attributes; and (2) a more abstract Surfaces Domain where "surface, corner, edge, shadow" are the entities; "convex, concave, behind, connected" are the relations; and "shape, tilt, albedo" are the attributes (see Sutherland, 1973). From a hierarchical view we might think of perception as an ordered sequence of unidirectional mappings from less abstract to more abstract representations and the differentiation of an intention as the successive mappings of the intention onto a series of progressively less abstract representations. But the argument from the coalitional and heterarchical interpretations of organization is that the conversation between abstract representations (domains, knowledge structures) is not one way. A fundamental result in artificial intelligence research on scene analysis is that while it is necessary to construct descriptions in many different domains, a procedure that exploits only unidirectional mapping from a lower domain to the next and higher domain is significantly limited in its capability to interpret a scene successfully (Sutherland, 1973). Success in scene interpretation is greatly enhanced by allowing a more flexible strategy in which processing in lower domains can use, as subprocedures, hypotheses generated about structures in higher domains (e.g., Falk, 1972).

Let us comment briefly on entities that in theory could be gathered together to form domains in action. On the basis of what has already been said, it would be logical for us to identify the entities in a representation with coordinative structures. In this regard it is important that reflexes can be arranged on a scale from complex and wide-ranging to simple and local. The organization of reflexes reveals "parallel hierarchies of complexity whose regularity and order leave little to be desired: local spinal reflexes, such as the flexion reflex, appear to be subsumed by reflexes requiring an intact spinal cord such as the scratch and long spinal reflexes, and these in turn are subsumed by pontine and medullary reflexes, such as the tonic neck and labyrinthine reflexes, and, at still higher levels, by locomotion and righting (Easton, 1972a: 593)." This suggests that we might equate the entities in *each* abstract representation of an act with coordinative structures, remarking that in higher domains an action plan is represented by functions defined over a relatively small number of large and complex coordinative structures and in lower domains by functions over a relatively large number of small and simple coordinative structures. We are thus provided with the following description of the evolving act: *an act evolves as the mapping by a heterarchically organized system of an intention onto successively larger collections of increasingly smaller and less complex coordinative structures, with each representation approximating more closely the desired action*.

There are many other ways we might conceivably characterize the entities of a representation in the building of a theory of action, but I hope the arguments that follow will pinpoint the special advantages conferred on a system in which the entities at all levels of representation are relatively autonomous structures. At all events, let us simply note at this juncture two contrasts between action as hierarchy and action as heterarchy. In one, the contrast is between the hierarchic strategy of a detached higher level dictatorially commanding lower levels and the heterarchic strategy of procedures constructing a representation in a higher domain entering into "negotiations" with lower domains in order to determine how the higher representation should be stated. In the other, the contrast is between the hierarchic principle of low-level structures unquestioningly responding to high-level instructions and the heterarchical principle of procedures establishing a representation in a lower domain reprocessing higher representations from the perspective of the special kinds of knowledge available to the lower domain.

The anatomical and neural structure of mechanisms related to movement suggests quite strongly that the fluidity called for by coalitional or heterarchical organization, the constant shuttling back and forth between domains, is not without basis. Consider, for example, the notion of internal feedback. Most generally the idea of feedback in behaving organisms is identified in two senses. In one sense, it is information that arises from the muscles as a direct consequence of their being active; in the other sense, it is information originating outside the organism as an indirect consequence of muscular contraction. The latter is often dubbed "knowledge of results." These senses of the concept of feedback are not exclusive, for they omit the afferent information that arises from structures within the nervous system in the course of an act's emergence. We refer to this feedback from the nervous system to itself as internal (Evarts et al., 1971) and it plays a central role in the evolution of coordinated activity.

In reference to internal feedback from spinal centers, Oscarson (1970) has remarked on the fact that a number of ascending pathways (at least six spinocerebellar tracts) are not especially well equipped to provide information about muscle contraction. Rather, the organization of these ascending paths suggests that they monitor activity in spinal motor centers, which in turn provide an abstracted account of the relation between themselves and other lower centers. This property of ascending paths fits the character of descending paths: most descending fibers terminate in interneuronal pools rather than passing directly to motor neurons. The basis for this arrangement may lie in the fact that the coordination of movement rests on the patterning of groups of motor neurons rather than on instructions to individual units, and the mapping between domains consists of predictions of how functional groupings of muscles (coordinative structures) will behave (cf. Arbib, 1972). Spinal centers thus provide a means for checking predictions against the current status of lower centers. Therefore, interneuronal pools may function as "correlation centers" (Arbib, 1972) reporting the degree to which an action plan is evolving as desired or indeed capable of evolving in the desired manner from a particular representation. At all events, there are probably many such internal feedback loops broadcasting the state of each level of the actor from executive to muscle (Taub and Berman, 1968), a highly desirable state of affairs from the perspective of a strategy in which executive procedures draw rough sketches and low-level procedures furnish needed details.

Our appreciation of the flexible relations between neuroanatomical structures supporting action is fostered further by recognition of the fact that signals from above can bias the abstracted accounts supplied by spinal centers. Many supraspinal mechanisms exert influences on the first synapse in ascending systems, i.e., the synapse between the peripheral afferent neuron and the second-order neuron which crosses the spinal cord to the tracts projecting to the brain (Ruch, 1965a). These influences from above are exerted mainly by motor areas and motor tracts, including the classically defined principal motor tract, the pyramidal (corticospinal) tract.

Current deliberations on the interrelations among the motor cortex, basal ganglia, and cerebellum may well be resolved on the acceptance of the coalitional formulation (see Kornhuber, 1974). We know that before the first signs of muscle innervation relevant to a particular movement significant changes occur in the activities of the cerebellum and basal ganglia, in addition to the motor cortex (Evarts et al., 1971; Evarts, 1973). This contrasts sharply with more traditional interpretations of basal ganglia-cerebellar processes operating as movement control and error-correcting devices coming into play only after the innervation of muscles. Rather, it would seem that these mechanisms gang together in the constructing and differentiating of action plans--they incorporate different procedures, each using the others as subprocedures as the situation demands. The structure of the cerebellum and its relations with other structures exemplifies the flexibility of neural computation in action. The cerebellum receives inputs from the entire cerebral cortex, projects to the motor cortex (Evarts and Thach, 1969), and is in two-way communication with the segmental apparatus of the spinal cord and thus with the structures that will actually execute the intended configuration of motions.

Thus the cerebellum can operate as a comparator, relating information about cerebral events to information about spinal events. The argument has been made that the cerebellum carries out a speeded-up differentiation of representations of the action plan, thereby providing a projection of their outcome and a basis for their modification. On this argument, the cerebellum plays a significant role in tailoring action plans to prevailing environmental and kinematic conditions prior to their realization as muscle events and thus prior to feedback from muscle contraction (see Eccles, 1969; Kornhuber, 1974).

EXECUTIVE IGNORANCE: EQUIVALENCE CLASSES AS INSTRUCTIONAL UNITS

Clearly, coalitional and heterarchical organization is far more flexible than hierarchical organization. Yet this flexibility is constrained in important ways. For example, in action there would be limits on the depth to which procedures constructing a representation on a higher domain may go in search of useful hypotheses. For any given higher abstract representation of an intention, the utility of knowledge about any lower domain would be inversely related to the degree of abstraction separating the two domains. Hypotheses about individual α-γ links (the smallest coordinative structures) regulating muscle contraction, for example, would not be useful to the determination of relevant large coordinative structures and related functions. And this, of course, is no more than a restatement of the degrees of freedom problem noted above. It then follows that while a representation of an intention in a higher domain is mapped into an immediately lower domain, the particular form that the representation will actually take in the lower domain cannot be known in advance, for the procedures operating in the lower domain have access to knowledge that is immaterial, in principle, to the procedures in the higher domain.

This form of "ignorance" has been duly recognized by students of action. We recall the earlier comment that the role of the executive (which is understood to be not a single neuroanatomical structure but a set of procedures engaging a number of neuroanatomical structures) is to modify the mode of interaction among elements at a lower level (Bernstein, 1967; Gelfand et al., 1971). As a general rule, however, it is argued that the executive does not have advance knowledge of which particular state, out of a set of possible states, a lower level will arrive at after a mode of interaction has been specified (Pyatetskii-Shapiro and Shik, 1964; Greene, 1971b). In this perspective, Greene (1971a:xxii) asks: "Can there be units of information that behave deterministically, even though the executive can rarely specify control functions more narrowly than to place them within broad classes of possible realizations?" Consider a situation in which the executive specifies a function transferring a given system into a "model" state. Now we may say that the "model" state serves not as a binding decree to be followed dogmatically by the system but rather as the identifier of a "ballpark," i.e., an equivalence class of states convertible into the "model" state. For the system, two states are defined as equivalent if they differ by a transformation that is realizable by the system. To Greene's (1971a, 1971b) way of thinking, the interconverting of states or functions is characteristic of low-level systems, so that a state or function specified by the executive (or for that matter any higher domain) may be substituted for by one from the same equivalence class but is more attuned to the current conditions operating within and around the system and to the system's privileged knowledge of the capabilities of other low-level structures. Similarly, executively specified functions determining the switching from one structure to another form another "ballpark," and low-level systems may autonomously interconvert transition functions of the same equivalence class as the need arises. By this reasoning the units of information that behave deterministically are not functions but equivalence classes of functions. [The reader should refer to

Greene (1971a, 1971b) for a more detailed and formal account of the various kinds of possible equivalence classes.]

In both of the above instances (specification of model states and transition functions) the executive instructions would be judged as satisfactory by the executive even though the instructions (functions) specified were not those actually carried out by the instructed systems. However, executive ignorance about which functions or states actually arise in the lower levels implies a high degree of uncertainty in executive commands, since for any given system the executive is specifying an unknown member of a family of possible functions or states. "This uncertainty introduces ambiguities and errors in an executive system's memory, commands, and communications to other executive systems" (Greene, 1971b:4-11). And we must expect these ambiguities and errors to be propagated through the action system during an act's evolution. The question arises, therefore, of how a configuration of motions is coordinated with precision and finesse. Indeed, in the face of this apparent chaos, we should ask how coordination can be achieved at all. We can only assume that the action system is so constructed and its procedures so related that these ambiguities and errors are immaterial to the differentiation of an action plan (cf. Pyatetskii-Shapiro and Shik, 1964).

For Greene (1971a; 1971b) the answer lies in the relations among the various equivalence classes: even though errors induced in one class may lead to erroneous specification in another, that specification would still be confined within the equivalence class of the desired function. Thus the equivalence classes as invariant units of information provide a means for specifying instructions in terms that are reliable and intelligible, even though an executive system is ignorant of the desultory character of low-level systems. We may summarize with Greene (1971a:xxiv-xxv): "Roughly speaking the equivalence classes serve as 'ballparks' into which it is sufficient for the executive to transfer the state: once the state enters the ballpark it will be automatically brought to the correct position without further attention--although ambiguities inevitably lead to erroneous signals, these signals will never be moved outside their correct ballparks or equivalence classes. Hence the equivalence classes seem to be systematically behaving units of information in situations in which the individual elements themselves will behave in haphazard fashion."[3]

---

[3] From what has already been said, it is evident that the derivation of a pattern of motions from its underlying representation is the cumulative result of the application of a long series of "rules." We should suppose, therefore, that there are regularities in the representation of an act in the highest domain that are obscured at the movement level by the application of these rules. In part, Greene's (1971a) equivalence classes are an attempt to recover the regularities, and the rules of the action system are defined by the conditions underlying the change in identity of functions and states, i.e., the interconverting of elements within a class. Obviously the enterprise undertaken by Greene has much in common with current approaches to problems in linguistics. Thus, in phonology, rules are sought that insert and delete and even change the segments specified in the underlying representation (Schane, 1973). It is perhaps important for the reader to treat Greene's ideas of procedures interconverting functions, states, and pieces (coordinative structures?) whose functions agree through some range (an equivalence class identified by Greene, which was not discussed above) as a comment on the kinds of neural computations performed in the course of transforming an action plan into innervational signals to muscles.

THE ACTION PLAN AND THE ENVIRONMENT

Taking stock of our analysis thus far, we may draw a rough sketch of how an action plan is represented in the highest domain, namely, as the specification of a subset of large coordinative structures that almost fits what is intended, and a set of functions on that subset (identifying the necessary equivalence classes) that will relate its elements both adjacently and successively in a particular way. Thus the serial nature of an act is said to arise not from the extemporaneous linking of component motions but rather from the differentiation of an already formed plan (cf. Lashley, 1951; Bernstein, 1967; Evarts et al., 1971; Pribram, 1971). We have not, however, made any comment on the relation between the action plan and perception. To rectify this omission, we return once again to the nature of an action concept, more precisely, A(x). We do so on the following rationale: if some common ground between the action concept A(x) and its perceptual counterpart can be identified, then perhaps we can gain some perspective on the relation between the action plan and the perceived environment in which the action plan is to unfold.

Consider, much as we did before, a sample of A's written by the same individual using different muscle combinations, e.g., one A may have been written by small motions of the fingers, another by large motions of a leg with the writing instrument grasped between the toes. Members of the sample will differ metrically: they will probably be of different sizes, of varying orientations, and of differing degrees of linearity, i.e., some will be written in curved strokes, while others will be virtually straight. And supposedly all members of the set will differ spatially in that they will occupy different locations on the page. On inspection, we would probably have little difficulty identifying each member as an instance of capital A. But in what sense are they equivalent? In geometry, figures are defined as equivalent with respect to a group of transformations. We say that two figures are equivalent if and only if the group contains a transformation that maps one figure onto the other. The group of transformations relevant to this discussion, i.e., relevant to our sample of A's, is clearly nonmetrical and, by elimination, must be topological. It is nonmetrical properties rather than metrical properties (which would be left undisturbed by the metric groups: the group of motions, the similarity group, and the equiareal group) that are of significance to the perceptual determination of membership in the class of capital A.

By the same token, the action concept supporting A-writing is determined by nonmetrical properties rather than metrical properties. After all, the sample of A's we are considering was the product of an actor, and the sample, as we have noted, is indifferent to metrics. Since this is no more than a paraphrase of an argument by Bernstein (1967), we should let Bernstein draw the relevant conclusion concerning the action concept of A: "The almost equal facility and accuracy with which all these variations can be performed is evidence for the fact that they are ultimately determined by one and the same higher directional engram in relation to which dimensions and position play a secondary role...the higher engram, which may be called the engram of a given topological class, is already structurally extremely far removed....from any resemblance whatsoever to the joint muscle schemata; it is extremely geometrical, representing a very abstract motor image of space" (p. 49).

In short, the action concept for writing A and the perception concept for identifying A share common ground in their dependence on nonmetrical properties, and it is not difficult to imagine an isomorphic relation between them (Turvey, 1974). But this is a very special case and we may ask: Does the action plan in general relate in similar fashion to the perceived environment? Bernstein (1967) hazards the guess that it does. For him the high-level abstract representation of an action plan may be construed as a projection of the environment relevant to the intention, where this projection relates to the environment topologically but not metrically.

Owing to the vagueness of this argument, we may feel that we have not really acquired any new insights (after all, what does it mean to talk about an isomorphism between action plans and environmental events?). Yet honoring our eccentricities for the present, we may acknowledge that we have reinforced our respect for the action plan and the action coalition/heterarchy. In earlier arguments, we established the fact that the high-level abstract representation of an action plan was not a projection of muscles and joints. In the current argument, we maintain that an action plan may be usefully construed as a projection of the environment. Therefore, we view the task of the action coalition/heterarchy as that of translating an abstract projection of the environment into joint-muscle schemata.

Some research by Evarts (1967) is of special relevance to these speculations. Evarts showed that when a monkey makes a movement of the wrist to counteract an opposing force (in a task in which the direction of force and the direction of displacement are varied orthogonally), recordings from unit cells in the motor cortex are related to the amount of force needed rather than to the degree of displacement. Moreover, this activity in the motor cortex is manifest prior to evidence of muscular contraction. As Pribram (1971) points out, Evarts's observation suggests that the representation at the motor cortex is a mirror image of the field of external forces. But by our account this "image" must represent the action plan at a fairly late stage in its differentiation and, in terms of the earlier analogy with linguistic theory, is more closely related to the surface structure of an act than to the deep structure. Indeed, Evarts (1973) has claimed recently that the representation of movement at the motor cortex, rather than identifying the highest level of motor integration (a classical point of view), is, on the contrary, much closer to the muscles and hence much lower in the organization of the action system than representations in other (traditionally lower) anatomical systems, such as the cerebellum and the basal ganglia.

Accepting the proximity of this motor cortical representation to the act's surface structure, we can see that by this level the action coalition/heterarchy has transformed a projection of topological properties of the environment into a projection of environmental contingencies (e.g., forces). According to Bates (cited by Evarts, 1973), force is the logical output for the motor cortex; velocity is the single integral of this quantity, and displacement is the double integral, and both of these quantities are theoretically more difficult to specify than force itself. Yet ultimately acts call for accurate displacements, and accurate displacements, in turn, call for a projection of metrical properties of the environment. We are led, therefore, by this reasoning to another description of the evolving act, namely, that the action plan unfolds as a series of progressively less abstract projections of the environment.

THE PROBLEM OF PRECISION AND THE CONCEPT OF TUNING

The realization of an action plan as a coordinated pattern of motions requires its translation from the crude language of coordinative structures to the precise language of muscles. Commerce between animal and environment reduces ultimately to the regulation of pairs of antagonistic muscle groups coupled together at joints. In the translation from abstract action plan to mechanical response, the α motor neuron stands as the penultimate component. The central question now is: How can α motor neurons specify to muscles the needed lengths and tensions when the terms length and tension are not in the lexicons of higher domains and hence, by definition, cannot be ingredients in an action recipe? In short, we seek to understand more fully the mechanisms through which the action system generates precise commands to muscles from crude commands to coordinative structures.

An instructive portrayal of the problem in limited form follows from the concluding comments of the preceding section--that a suitable output for the motor cortex is force. Suppose that subsequent processes, metaphorically speaking, integrate this quantity. Then, as already noted, the single integral will yield velocity and the double integral will yield displacement. But the particular displacement obtained for any given force will depend on the end-points or limits of integration. Thus specification of force alone is insufficient for the achievement of a desired velocity or displacement--the limits of integration must also be identified. How such "end-points" might be supplied in the relating of action to environment is the kernel of our problem.

In order to aid our inquiry, we now proceed to consider and illustrate properties of the spinal cord. Earlier we remarked that the role of higher levels of the nervous system is to pattern the interactions within and among coordinative structures. Let us now recognize that, in the main, coordinative structures have their origins in the spatially divided and relatively autonomous subsystems of the spinal cord. And let us modify our terms slightly to read: the role of the higher levels of the nervous system is to modulate interactions within and among neural mechanisms at the spinal level (cf. Obituary: Tsetlin, 1966; Gelfand et al., 1971).[4]

The segmental apparatus of the spinal cord is a functional entity well suited to the organization of coordinated activity. Its component structures are richly interconnected by a variety of horizontal and vertical linkages, providing an intrinsic system of complex interactions that is no less essential for the evolving act than supraspinal influences. The spinal cord is an active apparatus that does not passively reproduce instructions from above (Gurfinkel, Kots, Krinskiy, Pal'tsev, Feldman, Tsetlin, and Shik, 1971) and, indeed, may regulate its degree of subordination to supraspinal mechanisms (Sverdlov and Maksimova, 1965; Veber, Rodionov, and Shik, 1965). Several properties of spinal cord architecture and dynamics provide the basis for this interpretation. We take note of some of them here. First, of the great many interneurons in the spinal cord, relatively few are afferent neurons. Second, interneurons rather

---

[4] This point of view is also expressed by students of motor control in insects (e.g., Wiersma, 1962; Rowell, 1964; Weiss-Fogh, 1964).

than motor neurons are the terminal points for the majority of descending fibers from the brain. Third, the majority of synapses in the spinal cord are formed by connections between spinal neurons, and relatively few are formed from axons coming from the brain and spinal ganglia. And fourth, reciprocal facilitation and inhibition, and myotatic reflex action are all processes at the segmental level (see Gurfinkel et al., 1971).

The integrity of the spinal cord rests on the fundamental servoprocess manifest in the α-γ link regulating muscle contraction. On this servoprocess are built the intra- and intersegmental reflexes. We suppose that the modus operandi for integrating reflexes (the basis of the set of coordinative structures) in coordinated action exploits rather than disrupts the fundamental servomechanism. Indeed, this will prove to be the key notion for unraveling the problem of how precise instructions are formulated at the level of α-motorneuron activity.

But before examining the evidence for this view, we observe that in the unfolding of the action plan on the segmental apparatus, the responsibility for demarcating coordinative structures and for the parsing of those structures may devolve on separate neuroanatomical systems. Greene (1971b) cites a series of experiments by Goldberger in which the corticospinal and brain-stem spinal paths in monkeys were interrupted. With corticospinal interruption the animal can no longer inhibit unwanted components of a coordinative structure such as the group of muscle contractions that extend joints of the same limb. Thus, for example, when presented with food that he must stretch for, the monkey reaches out with extended limb but cannot then close his fingers to grasp the food. If with the joints flexed the animal grasps food placed close to him and raises it to his mouth, he cannot then let go. In contrast, brain-stem spinal interruption appears to impede the animal's ability to restrict the evoked coordinative structures to those relevant to the task. When extending an arm to reach for food at a distance, the group of contractions that rotate the limb, or those that raise the limb, may come into play in addition to the task-relevant group of limb extensors.

In short, we see that the delimiting of coordinative structures and the manner of their decomposition are effected in the segmental apparatus by instructions from separate mechanisms. But now we must pass from this gross differentiation of the action plan at the segmental level to the finer differentiation afforded by the fundamental servomechanism (or, more aptly, the fundamental coordinative structure).

The main body of a muscle consists of extrafusal fibers that on contraction alter the relative positions of the bones to which they are attached. The innervation of extrafusal fibers is supplied by α motor neurons. Within the main body of the muscle are intrafusal fibers that are wrapped around the middle by the terminals of sensory fibers. These sensory fibers and the intrafusal muscle fibers to which they attach are referred to collectively as a muscle spindle. Muscle spindles connect to the extrafusal fibers at one end and to a tendon at the other and are therefore "in parallel" with the extrafusal fibers. Two functionally distinct spindle components can be identified: a static component that is sensitive to the instantaneous muscle length and a dynamic component that is sensitive to the rate of change of muscle length (Matthews, 1964). On contraction of the intrafusal fibers, the spindle receptors register the difference in length and the difference in rate of change of length between the intrafusal and extrafusal fibers. The induced receptor excitation is communicated to the linked α motor neurons, which respond by recruiting more extrafusal fibers until the discrepancies in length and velocity have been annulled. Thus, in a situation in which a load is applied to a muscle extending it beyond its resting length, the spindle feedback provides an autonomous means of tailoring the muscle response to the new conditions. This negative feedback system identifies the fundamental servomechanism; it is now incumbent upon us to show that this servomechanism is biasable in ways of considerable importance to the theory of action.

Intrafusal fibers, like extrafusal fibers, have a source of innervation, the γ motor neurons. These motor neurons fall into two relatively independent classes, the γ static and the γ dynamic, regulating, respectively, the static and dynamic components of muscle spindles (Matthews, 1964). Again, γ motor neurons, like α motor neurons, are under high-level control, but the motor nerves that project from brain to γ motor neurons and those that project from brain to α motor neurons are largely separate, and thus it is optional whether the spindles and the main body of a muscle contract and relax together. "The spindles could therefore be activated while the main muscle remained passive, and vice versa" (Merton, 1973:37).

We see, therefore, that the γ system allows for the modulation of the fundamental servomechanism. The γ-static motor neurons can control the equilibrium state of the servomechanism, while the γ-dynamic motor neurons can control the "damping" of the servomechanism, i.e., the rate at which it achieves equilibrium. Thus, the servomechanism is not only informed of what it has done, but more importantly it can be informed of what it must do.

This completes the elementary description of the biasable nature of the fundamental servomechanism; but one or two points remain to be considered.

In addition to the biasable feedback loop signaling length and velocity through spindle receptors, there is another that signals muscular force through tendon organs. The signals conveying force feedback converge on interneurons on their way to α motor neurons. As before, interneurons can be manipulated by higher-level instructions so that the inhibitory effects of force feedback on α-motor-neuron activity can be modulated. The biasable feedback loops conveying length, velocity, and force are inextricably linked in the regulation of the servomechanism. So we may expand on our comments above. While higher-level control signals to α motor neurons set the servomechanism going, the higher-level control signals to the γ system and to the interneurons transmitting force information from the tendon organs function less as instigators of movement than as modulators of the gain of the feedback loops, that is to say, they serve to adjust the ratios of the outputs of feedback loops to their inputs. This principle of higher-level modulation of spinal reflexes is generalized to the segmental mechanism of reciprocal inhibition by which we understand that spindle activity not only impinges on an α motor neuron of its own muscle but also, via inhibitory motor neurons, on an α motor neuron of the antagonist muscle. Spindle output contributes both to agonist contraction and to antagonist relaxation in the regulation of pairs of muscles controlling a joint. Clearly, from all that has been said, the reflex interplay between agonist and antagonist is biasable.

We are now in a position to identify the property of the spinal cord that is central to our current concerns (and for which we shall shortly provide evidence), namely, that the system of segmental interactions is biasable and as a general strategy the activating of coordinative structures occurs against a background of spinal mechanisms already prejudiced toward executive intentions. Thus it may be argued that the control of movement is in many respects the reorganization or tuning of the system of segmental interactions and that this attunement precedes the transmission of activating instructions to coordinative structures (Gelfand et al., 1971; Gurfinkel et al., 1971).

Before we pass from this elementary discussion of tuning to take up the topic in earnest, let us glance at two examples of how α-γ linkage might embellish a relatively simple instruction. For the first, consider the previously discussed action of stepping, recognizing the fact that a double-joint flexor of the hip also produces extension at the knee. The high-level representation of stepping can be said to specify crudely a general "flexor plan" for the limb (Lundberg, 1969). The knee flexors innervated on this plan are strong enough in the early phases of the movement to prevent extension, but as flexion proceeds, the double-joint muscle will become stretched, inducing an intense discharge of spindle activity. The spindle feedback will impinge upon hip-flexor α motor neurons and also produce reciprocal inhibition of the motor units belonging to knee flexors. Thus, during the swing, the knee flexors originally innervated on the flexor plan will suffer inhibition from the spindle activity of the double-joint muscle. The upshot of this interplay is the differentiation of the broadly stated flexor plan into coordinate stepping (Lundberg, 1969).

For the second example, consider a sudden change in the loading on an outstretched arm, e.g., a heavy object is placed into the hand, in a task where the arm's inclination to the ground is to be kept constant. We may suppose that instructions to the appropriate coordinative structures quickly bring the arm back into a close approximation to the desired position, with spindle feedback coming into play to finely tune the terminal point of the trajectory (cf. Navas and Stark, 1968; Arbib, 1972) and to maintain the arm in its desired position under the new conditions. In this sense, we construe the α and γ systems as participating in a "mixed ballistic-tracking strategy" (Arbib, 1972:134), with the α system determining the ballistic component that gets the limb quickly into the right ballpark, and with the γ system determining the superimposed tracking component that supplies the needed refinements.

From what has been said about the segmental apparatus and its pretuning, we understand that the α-γ processes of the two examples cited above do not take place in a vacuum. Rather, they occur against a backdrop suitably colored by supraspinal influences. It can be demonstrated that in the final 30 msec or so preceding a movement there is a pronounced enhancement of the effect of reciprocal inhibition on the future antagonist of that movement (Kots and Zhukov, 1971, see below). What this suggests is that a motion like stepping or raising the arm is anticipated through the supraspinal tuning of the segmental mechanisms of reciprocal inhibition. We should also recognize that the backdrop for a voluntary act is not limited to adjustments in spinal reflexes. Thus, for example, when an arm is raised voluntarily by a person standing upright, it is possible to observe in the period immediately prior to the first signs of arm muscle activity, anticipatory activity in a number of muscles of the lower limb and trunk (Belen'kii, Gurfinkel, and Pal'tsev, 1967). Figuratively speaking, when one moves an arm in a standing position, one first performs "movements" with the legs and the trunk and only then with the arm.

In summary, we have considered in preliminary but sufficient fashion the kinds of mechanisms that effect significant variation in the behavior of low-level holons without infringing on their autonomy. These mechanisms suitably controlled from higher domains allow for the precise regulation of muscular contraction. The gist of the whole matter is given in a short paragraph by Pribram (1971:225):

> When reflexes become integrated by central nervous system activity into more complex movements, integration cannot be effected by sending patterns of signals directly and exclusively to contractile muscles, playing on them as if they were a keyboard. Such signals would only disrupt the servoprocess. In order to prevent disruption, patterns of signals must be transmitted to the muscle receptors, either exclusively or in concert with those reaching muscle fibers directly. Integrated movement is thus largely dependent on changing the bias, the setting of muscle receptors.

To this we need only add that the modulation of interneuronal pools must also play a significant part in the biasing of servomechanisms.

MOVEMENT-RELATED SEGMENTAL PRETUNING

Our task now is to view evidence for segmental pretuning in voluntary movements. But before we do so, we must acknowledge that the available evidence comes from experiments in which the performers on cue are required to flex a knee, bend an elbow, extend a foot, and, in general, to execute movements whose trajectories, velocities, degrees of displacement, and so forth, are indifferent to the environment. To my knowledge, there are no experiments on segmental pretuning for acts that depend on the detection of environmental properties for their performance. As a precautionary measure, therefore, we shall distinguish between movement-related and environment-related pretuning. The former will refer to the changes in the segmental apparatus preceding the execution of a simple voluntary motion that is unrelated to the environment, in the sense that neither the actor's position with respect to the environment nor the position or orientation of objects with respect to each other and to the actor are altered by the motion. The latter, environment-related pretuning, will refer to segmental changes that precede actions or, more precisely, components of actions that are environmentally projected, in that their purpose is to displace the actor with respect to the environment or to displace (or rotate, or reflect) objects with respect to the actor, or both.

The goal we are approaching slowly is that of roughly and approximately understanding how seeing enters into doing. To this end, we shall need to extrapolate from movement-related pretuning to a general picture of how environmental properties control action. At all events, our immediate concern is with movement-related pretuning, and we begin with some general comments.

The methodology for investigating movement-related pretuning has much in common with the methodology that characterizes the information-processing approach to visual perception (cf. Haber, 1969), which seeks to determine how visual "information" is modified in the course of its flow in the nervous system. Thus techniques of masking, delayed partial-sampling, and reaction time are used to assess the correlation between stimulus and response at varying delays after visual stimulation.

In a similar if less sophisticated vein, the tuning experiments we shall consider in this section judiciously apply the principle of probing the nervous system, in particular the spinal cord, in the interval elapsing between a warning signal and a cue to respond, or (more specifically) between a cue to respond and the first signs of activity in the agonists executing the motion. The probes are simple reflexes elicited during the interval, with the latency and amplitude of the reflexes (recorded by the electrical response of the corresponding muscles) taken as indicants of the state of the segmental apparatus prior to the movement. The reflexes used for this purpose have generally been tendon reflexes elicited by a tap and the Hoffman or H-reflex, which is a monosynaptic reflex in the gastrocnemius-soleus muscle group and elicited by electrical stimulation of the tibial nerve in the popliteal fossa (Hoffman, 1922).

As an introduction to the procedure and to the observations that will be of interest to us, we consider two exemplary experiments. In the first (Gurfinkel et al., 1971), the participant is seated with legs flexed and on command extends one leg, the responding leg remaining constant throughout the experiment. Surface electromyographic (EMG) recording of activity in the quadriceps femoris muscle reveals that the tibia extension occurs at a latency of 160 to 180 msec. If the patellar reflex is evoked in the same leg within 100 msec or so of the cue to respond, the amplitude of the reflex is unaffected (compared to the control condition in which no command to extend is given). However, if the patellar reflex is elicited beyond this period, then its amplitude is enhanced the closer to the command that it is elicited. We infer, therefore, that the state of the segmental apparatus has been altered prior to activation of the muscle group extending the knee. In the second example (Gottlieb, Agarwal, and Stark, 1970), the participant is again seated normally but with the right leg extended, knee slightly flexed, and the foot firmly strapped to a plate to which he transmits an isometric force through either plantar flexion or dorsiflexion. The task of the participant is to match the level of his foot torque with a target level specified by the experimenter in what is, essentially, a continuous tracking task in which the varying target level of torque and the level of the participant's matching torque are displayed on a scope viewed by the participant. The H-reflex is elicited at different delays subsequent to the target adopting a new level, and both the H-reflex and the activity in the gastrocnemius-soleus muscle group (agonist in plantar flexion) and the anterior tibial muscle (agonist in dorsiflexion) are measured. Summarized briefly, the results are that the amplitude of the H-reflex is distinctly augmented if elicited in the period 60 msec prior to the initial signs of voluntary motor unit activation in plantar flexion, and is generally inhibited in approximately the same interval before the first signs of voluntary dorsiflexion. Again we infer that there are changes in the spinal cord that precede agonist activation.

How specific is segmental pretuning? Consider initially some further experiments reported by Gurfinkel et al. (1971). In one, two movements are executed by the subject on separate occasions: flexing the leg at the hip joint and extending the knee. Measures are taken of the tendon reflex of the rectus head of the quadriceps femoris muscle, which spans two joints, and of the lateral head of the same muscle, which spans only one joint. When the hip is flexed, a premovement increase is observed only in the amplitude of the tendon reflex elicited by stimulation of the rectus head. But it is important to note here that while the rectus head of the quadriceps femoris is involved in hip flexion, the lateral head is not. By contrast, when extension of the knee is called for, a movement that involves both heads of the quadriceps femoris, both reflexes are significantly increased in amplitude prior to signs of voluntary motor neuron activity. In another experiment, Gurfinkel et al. observed that in the 70 to 80 msec prior to flexing the leg at the knee, the patellar reflex is amplified, but they also observed that the patellar reflex is amplified (although not to the same degree) prior to flexing the elbow of the arm ipsilateral to the leg in which the patellar reflex is elicited. Therefore, we may conclude that both specific and nonspecific changes in the segmental apparatus of the spinal cord precede voluntary motor unit activation.

This conclusion is buttressed by some other experiments that examine the changes in the spinal cord during the period intervening between a warning signal and the signal to execute a given movement. We take, as examples, experiments reported by Requin and Paillard (1971) and by Requin, Bonnet, and Granjon (1968). In these experiments the movement is extension of the foot (plantar flexion), and both tendon- and H-reflexes are recorded from both the participating leg and the nonparticipating leg. Following the warning signal, there is evidence of an increase in the amplitude of the reflexes measured in both legs, an increase that persists for the reflexes of the nonparticipating leg but that is progressively depressed in the participating leg with greater proximity to the cue to respond. In short, these experiments provide evidence that after a warning signal and before a signal to execute a movement, there occur both a nonspecific change in spinal sensitivity and a specific change related to the motor neuron pool--the servomechanisms--about to be involved in the forthcoming movement. In the view of Requin and his colleagues, the depression of the reflex amplitude in the participating leg is due, under the conditions of the warning signal procedure, to central (supraspinal) influences that selectively "protect" the direct participants in the movement from irrelevant influences exerted upon them prior to their activation.

Evidently, in the premovement period, specific changes occur in the feedback loops related directly to agonist regulation. But can we demonstrate similar effects in the more extended feedback loop relating the state of an agonist to its antagonist? Two experiments by Kots (1969a) and Kots and Zhukov (1971) provide an answer.

Kots (1969a) wanted to know whether the enhancement in the H-reflex excitability of the motor neurons of the gastrocnemius evidenced in the latent period of a voluntary movement depended on the role the gastrocnemius was to play in the forthcoming movement. To this end, the H-reflex amplitude was measured in the gastrocnemius when it was the future agonist of the movement, i.e., in plantar flexion, and when it was the future antagonist of the movement, i.e., in dorsiflexion. It was observed that following a command to move, the amplitude of the H-reflex was significantly enhanced in the period beginning 60 msec prior to the first signs of voluntary motor unit activation only when the gastrocnemius was future agonist. When the antagonist role was assumed, the H-reflex was neither enhanced nor depressed in the latent period but was found to decline sharply immediately following the first myogram signs of motor unit activity in the agonist muscle, the anterior tibial.

It would appear, therefore, that the effect of tuning was specific to agonistic activity and that the failure to detect a depression in the H-reflex when the gastrocnemius was the antagonist suggests that the "positive" priming of agonist centers is not paralleled by a "negative" priming of antagonist centers.

The absence of antagonist depression in Kots's (1969a) experiment contrasts with the evidence of such depression in the experiment of Gottlieb et al. (1970), described above. While not stating it explicitly, the report of the latter experiment implies that plantar flexion and dorsiflexion were mixed randomly in the course of an experimental session; in Kots's experiment, the two response modes were examined separately, and this difference in procedure may account for the difference in results. At all events, taken together, the two investigations suggest that depressing the motor neurons of antagonistic muscles in the latent period of voluntary movement is not a necessary concomitant of tuning agonists. However, as we shall see in the second experiment of Kots and Zhukov (1971), there is indeed an adjustment made in the inhibitory influences on the antagonists of a movement during the latent period that is not manifest as a depression in motor neuron excitability.

The Kots and Zhukov (1971) experiment made use of paired stimulation, a procedure that is comparable to the forward masking procedure commonly used in visual information-processing experiments (e.g., Turvey, 1973); essentially, a leading stimulus is used to impede the response to a lagging stimulus. For Kots and Zhukov the leading member of the stimulus pair was electrical stimulation of the peroneal nerve and the lagging member was electrical stimulation of the tibial nerve. Peroneal nerve stimulation elicits a direct response (the M-response) in the motor neurons of the anterior tibial muscle without accompanying M- and H-responses in the gastrocnemius-soleus muscle group. Tibial nerve stimulation, as we have already seen, elicits the monosynaptic H-reflex of the gastrocnemius-soleus group. The H-response in the gastrocnemius-soleus group is significantly depressed when elicited very shortly after peroneal stimulation, say 2 to 4 msec. The brief latency of this effect implies that it is realized by the "spinal apparatus of reciprocal inhibition" (Kots and Zhukov, 1971). Therefore, we can exploit this paired-stimulation procedure to monitor the state of reciprocal inhibition mechanisms during the latent period of a voluntary movement. Thus Kots and Zhukov sought to determine whether the impairment in the H-reflex induced by prior peroneal stimulation was intensified in the latent period of voluntary dorsiflexion. In more general terms, they sought to determine whether there is pretuning of the mechanisms of reciprocal inhibition. During dorsiflexion, reciprocal inhibition would protect the anterior tibial muscle from the antagonistic response of the gastrocnemius-soleus muscles; Kots and Zhukov looked to see if this mechanism was primed for its task before voluntary activation of the anterior tibial motor neurons. The experiment showed that in the final 30 msec prior to dorsiflexion, the paired-stimulation effect was significantly enhanced and, moreover, that this enhancement could not be due to a reduction in excitability of the motor neurons of the gastrocnemius-soleus group, since the H-response in the absence of preceding peroneal stimulation was unaltered during the latent period. Of course, this is what Kots (1969a) had found before.

Collectively, the experiments we have discussed suggest that a profound reorganization of the spinal cord precedes movement.[5] There are both nonspecific

---

[5] While we have chosen to discuss only experiments using simple, single movements, we should note that other experiments have examined patterns of segmental pretuning in the performance of sequential and rhythmic movements (e.g., Kots, 1969b; Surguladze, 1972). Thus Kots (1969b) showed that in the sequential performance of two movements opposite in direction in the ankle joint, the segmental organization for the second movement is realized during the execution of the first.

and specific components of this reorganization, and the latter have been shown to include the mechanisms of reciprocal inhibition in addition to the servomechanisms regulating agonist activity. Moreover, the reorganization of the interaction among neural mechanisms at the spinal level follows the pattern of initially diffuse becoming more localized the closer in time to the manifestation of the desired movement (Gelfand et al., 1971). It is evident, therefore, that in the differentiation of an action plan the realization of instructions to coordinative structures is determined by the state of these structures on receipt of the instructions (cf. Gurfinkel and Pal'tsev, 1965).[6] Since the argument is that coordinative structures when activated perform in a relatively autonomous fashion, it follows that the details of their performance are very much determined by the state of the segmental apparatus at the time of activation.

TUNING AS PARAMETER SPECIFICATION

The elegance of tuning as a control process is that it permits the regulation of a system without disrupting the system's autonomy. So far we have illustrated this principle only in the control of servomechanisms at the level of muscle contractions. But this should not blind us to the likelihood of tuning as a general principle fundamental to all domains of the action coalition/heterarchy. We recall the comment that actions are produced by fitting together substructures, each of which deals relatively autonomously with a limited aspect of the action problem. In addition, we recall that each domain may be construed as a representation in which relations are defined on a set of autonomous structures, with the size of these structures becoming progressively smaller and their number progressively larger as the action plan is mapped into progressively less abstract representations. We now hypothesize that into each representation tuning functions may enter as modulators of coordinative structures.

Miller, Galanter, and Pribram (1960), discussing the acquisition of the skill of typing, suggest that the student typist learns to put feedback loops around larger and larger segments of her behavior. We might well suppose that this notion applies beyond typing to skilled acts in general, and with internal feedback loops in addition to the more commonly understood forms of feedback. We can imagine tuning functions of a more abstract kind related to the modulation of feedback between action segments that collectively behave as relatively autonomous units in the performance of any given skilled behavior. Again, tuning would permit appropriate variation without disrupting, in these instances, the acquired self-regulating procedures.

---

[6] To demonstrate this point, Gurfinkel and Pal'tsev examined the effect of eliciting a tendon reflex subsequent to a cue for voluntary movement (extension of the knee) on the latency of the voluntary movement. They found that the latent period of voluntary extension was linearly dependent on the time at which the reflex was elicited: the later the reflex was elicited, the longer the latency of the voluntary movement. In addition, they showed that this effect held even when the reflex was elicited in the leg contralateral to that executing the voluntary leg extension. It is assumed that the reflex induces a change in the segmental apparatus and that the realization of commands for leg extension is therefore dependent on the prevailing state of the system of spinal relations. Gurfinkel and Pal'tsev suggest that the basis for this effect is adjustments in the states of interneuronal pools.

To extend our usage of tuning, we shall adopt a most important and provocative hypothesis, namely, that *tunings parameterize the equivalence classes of functions specified by executive procedures*. This follows from Greene's (1971a, 1971b) contention that the smallest units of information available to an executive are probably not functions but families of functions parameterized by possible tunings.

Although we are here attempting to pass beyond the idea of tuning limited to the fundamental servomechanism, we may profitably exploit our earlier discussion of that mechanism to illustrate the notion of tuning as parameter specification. We take as our departure point the experimental and mathematical analysis supplied by Asatryan and Fel'dman (1965) and Fel'dman (1966a, 1966b) of the maintenance of joint posture and of the simple voluntary movement needed to achieve a desired angle of joint articulation. Consider a simple mass-spring system defined by the equation $F = -S_o(\underline{l}-\lambda_o)$, where F is the force, S_o is the stiffness of the spring, $\underline{l}$ is the length of the spring, and λ_o is the steady-state length of the spring, i.e., the length at which the force developed by the spring is zero. This simple mass-spring system is controllable to the degree that the parameters S_o and λ_o are adjustable. Changing λ_o with S_o constant generates a set of nonintersecting characteristic functions, $F(\underline{l}) = -S_o(\underline{l}-\lambda)$, and changing both parameters generates a set of functions, $F(\underline{l}) = -S(\underline{l}-\lambda)$, that will pass through all points in the plane defined by the cartesian product $F \times \underline{l}$.

Let us now suppose that a joint-muscle system is analogous to our simple mass-spring system. In this case, we can argue that the problem of controlling a joint-muscle system reduces to that of fixing certain characteristics of the system, i.e., of setting mechanical parameters of the muscles, or more precisely, of setting biases on the fundamental servomechanisms. In the analogy, the characteristic functions of a joint-muscle system are of the form $M(\phi)$, where M is the total muscular moment and ϕ is the joint angle. And each $M(\phi)$, therefore, is determined by the mechanical parameters of the muscles regulating the joint: $\lambda = (\lambda_1,\lambda_2,\ldots\lambda_n)$, $S = (S_1,S_2,\ldots S_n)$, where n is the number of muscles.

Given the foregoing comments, let us now consider experiments using the technique of partial unloading--a technique (which we will shortly describe) that would assay the characteristic properties of an ordinary spring. Asatryan and Fel'dman (1965) sought to demonstrate that for a given situation the variations of muscular moment as a function of joint angle (or vice versa) are defined by an initial setting of the parameters, i.e., by a characteristic function. For purposes of analysis, we shall refer to the state of the joint-muscle system as α, where α is defined by the vector (M,ϕ). When M and ϕ are constant for some period of time, then α is a steady state of the system.

The experimental methodology may be described briefly. The participant's forearm is fixed on a horizontal platform whose axis of rotation coincides with the axis of flexion and extension of the forearm. The horizontal platform is attached to a simple pulley system supporting a set of weights that can be selectively unloaded. At the outset of a trial, the participant establishes a steady state, α_s, of the joint-muscle system: given a specified angle of articulation, the participant must establish a muscular moment to compensate for the effect of the moment of external forces--determined by the weights and their direction of pull on the horizontal platform--opposing flexion (or extension) of the joint. Thus, for a standard initial opposing force, different steady states,

α_s, can be established for different joint angles, ϕ_s. Once the steady state is established at a given ϕ_s, the participant is then asked to close his eyes and the weight is unloaded, with the amount of unloading varying across trials. The new angle of articulation--the new steady-state α_s'--to which the arm briefly moves following unloading (and before the participant can make compensatory adjustments) is recorded. From a series of experiments such as the one we have described, Asatryan and Fel'dman (1965) demonstrated that for all possible initial states α_s of the joint-muscle system, a set of nonintersecting functions, $M_s(\phi)$ [or $\phi_s(M)$], are generated relating muscular moments to the new steady-state angles of the joint. Moreover, they showed that the form of the function $M_s(\phi)$ does not depend on the external moments but is determined unambiguously by the parameters of the initial state of the system. (This was demonstrated by using a set of external moments that were rising functions of joint angles and a set that were diminishing functions of joint angles.) So we conclude that the function $M_s(\phi)$ for each α_s is an invariant characteristic of the joint-muscle system: if the system is perturbed, it will follow a trajectory of states leading to a new state of equilibrium, where both the trajectory and the equilibrium state are defined by the parameters fixed in the initial steady-state α_s. And since the curves are nonintersecting, the transition from one $M(\phi)$ to another requires changing λ_i with little but preferably no change in S_i (Asatryan and Fel'dman, 1965). It would seem, therefore, that a joint-muscle system does behave like a spring, i.e., like a vibratory system, and that the action structures can choose parameters for this "spring" in accordance with the prevailing conditions. For a brief period of time following perturbation, until new parameters of the spring can be specified, the joint-muscle system behaves in the way we would expect the chosen "spring" to behave.

We now proceed to develop this theme through the experiments of Fel'dman (1966b). These were conducted with a slight variation on the apparatus described above. The pulley-weight system was replaced by a detachable spring that opposes flexion of the joint but is insufficiently taut to prevent flexion. At rest, the joint is flexed at an angle ϕ_o, and on the occurrence of an auditory cue the participant must establish as rapidly as possible and without the aid of vision the steady angle ϕ_1. (The participant is given a practice session so that he can achieve ϕ_1 with a minimum of error.) During a series of trials, the spring is occasionally detached within the period subsequent to the auditory cue and prior to movement. Now suppose that at the outset of a trial a fixed invariant characteristic $M(\phi)$ has been determined for the attainment of a steady-state α_1, corresponding to the desired angle ϕ_1. In the steady-state α_1, $\phi = \phi_1$ and $M = M_e$, where M_e is the moment of force provided by the spring attached to the platform. But when the spring is detached, a new steady-state $\alpha_1' = (0,\phi_1)$ is required to achieve the same angle of articulation ϕ_1, which means, of course, that a new invariant characteristic $M'(\phi)$ is needed. The question is: Can the transition from one invariant characteristic to another be effected during the execution of the movement? If it cannot, then when the spring is detached, the joint will move to the angle ϕ_2 determined by the characteristic function $M(\phi)$. In the space (M,ϕ), ϕ_2 will be at the intersection of $M(\phi)$ with $M = 0$ (since $M_e = 0$). The results of the experiment reveal that during the rapid establishment of a desired steady angle in the joint, correction of the invariant characteristics of the joint-muscle system (correction of the parameters defining the projected steady state) does not occur. The correction is made only after the achievement of the new steady state (corresponding to ϕ_2) when the error becomes obvious.

Now we wish to prove that the joint-muscle system truly behaves in this situation like a mass-spring system; although the movement of such a system as a whole is determined by the initial conditions, the equilibrium position does not depend on them and is determined only by the parameters of the spring and the size of the load. Using the paradigm described above, we attach a pulley-weight system opposing extension such that release of the weight induces passive extension in the joint. Thus there are two external moments operating on the limb: a spring-opposing flexion and a weight-opposing extension. The participant becomes acquainted with the situation in which the spring is detached in the latent period before movement to the intended angle ϕ_1, bringing about passive flexion of the joint. But on some occasions the weight is also detached, leading additionally to passive extension before the voluntary movement. The results show that these rather radical and unpredictable changes in the initial conditions do not alter the behavior of the joint-muscle system: the trajectory of the system is still determined by the initial setting of parameters, i.e., it moves to the state defined by the characteristic function $M(\phi)$ established at the outset of the trial. In brief, the equilibrium position is independent of the initial conditions (Fel'dman, 1966b).

In further analyses, this time of rhythmic movements of the joints, Fel'dman (1966b) was able to demonstrate that there is an independent parameter setting for the dynamics of the joint-muscle system. Therefore, we may envisage the set of fundamental servomechanisms (the α-γ links together with the tendon feedback loops) regulating joint flexion and extension as collected together into a single vibratory system for which "static" and "dynamic" parameters can be specified. Choice of static parameters for the system determines the aim of a movement (the final steady state) independently of initial conditions; choice of the dynamic parameters determines (to a large extent) the rate and acceleration of the movement and also its form (aperiodic, oscillatory, etc.) (Fel'dman, 1966b).

This analogy between systems controlling action and vibratory systems suggests that we may usefully conceive of coordinative structures in general as biasable, self-regulating vibratory systems. In their simplest forms, such systems might be modeled by the following second-order homogeneous linear differential equation with constant coefficients:

$$\underline{m}\, X''(t) + \underline{k}\, X'(t) + \underline{s}\, X(t) = 0$$

where $X(t)$ is the function relating the displacement of the system from a steady state to time.[7] In such a system the setting of the parameter $\underline{s}$ defines the "stiffness" of the system and thus its equilibrium state, and the setting of the parameter $\underline{k}$ defines the friction or damping constant that determines the rate at which the system achieves equilibrium and the form of its behavior, i.e., whether

---

[7] This simple linear differential equation is given only to illustrate a principle. It is not meant to model (although it might) an actual coordinative structure. If we were to make the illustration more realistic and more general, we would need to consider forced vibration in addition to free vibration, and to concern ourselves with equations in which the applied force varied with time or acted in an arbitrarily short interval.

it oscillates or not. By way of summary, we have seen that the functional tuning of the segmental apparatus of the spinal cord may be likened to the specification of the parameters s and k for vibratory systems. On the assumption that all coordinative structures behave as vibratory systems, then tuning as parameter specification emerges as a viable procedure for adjusting the behavior of selected coordinative structures at all levels of abstraction of the action coalition/heterarchy. Thus, while some coordinative structures autonomously coordinate a greater number of pieces of the action apparatus than other coordinative structures (compare, for example, two classes of basic coordinative structures, the long spinal reflexes and the flexion reflexes), the manner of their attunement is fundamentally the same.

We now address the important question of whether the tuning and activation of autonomous systems are governed by the same mechanisms. Again, we shall proceed on the assumption that the regulatory principles for large systems follow very much the pattern of small systems. This permits us the latitude of extrapolating from the tuning of small systems, e.g., the fundamental servomechanism, about which we know something, to large systems, about which we know very little. The evidence of segmental pretuning suggests, among other things, that the nervous system has available a means of selectively raising and lowering the gain of spindle and tendon organ feedback loops. Indeed, the comment was made earlier that the control of the α and γ systems is largely separate, so that it is optional whether or not the two systems are concurrently active. But in the experiments we have taken as evidence for segmental pretuning, can a case be made for the selective modulation of servoprocesses independent of instructions sent specifically to activate α motor neurons, either directly or indirectly through γ motor neurons? In experiments exploiting the H-reflex and plantar flexion, such as those of Gottlieb et al. (1970), we might suppose that changes in the reflex during the latent period reflect nothing more than the increasing excitability of gastrocnemius-soleus motor units brought about by direct supraspinal signals to the α motor neurons. Or, in a similar vein, the increase in the H-reflex represents the increased excitability in the α-motor-neuron pool of the gastrocnemius-soleus group in response to stimulation from the γ system, which is in turn responding to directions from above. In these accounts, the variation in the reflex is not an independent event but an epiphenomenon of α-system innervation; that is to say, the voluntary electromyographic (EMG) and the H-wave variations are manifestations of the same controlling input. Against this argument, however, Gottlieb et al. (1970) point out that changes in the waveform and amplitude of the H-reflex are not correlated with changes in the agonist or antagonist EMG and, in addition, that the time courses of the recordings are clearly different. From their point of view, it is much simpler to propose that for their particular form of voluntary movement, there is a means for modulating the H-reflex (and by inference, the fundamental servomechanism) that is separate from the means for activating α motor neurons. In more general terms, we may conjecture that the mechanisms of tuning and activating coordinative structures are largely separate.

THE RELATIONSHIP BETWEEN THE EXECUTIVE AND TUNING

Let us summarize briefly our thinking thus far. The executively specified action plan identifies the relevant subset of coordinative structures and a set of functions on that subset (identifying the necessary equivalence classes) that will modulate its elements and relate them in a certain fashion. In the course of spelling out the action plan through successive procedures within the action coalition/heterarchy, the functions identified by the executive may be substituted for by functions more suited to the current low-level conditions of the system. The interconverting of functions, however, leaves the equivalence classes invariant. Of these interconversions and of the low-level realization of the details of the action plan, the executive remains virtually ignorant.

The eventual activation of coordinative structures takes place against a background of prearranged interactions within the segmental apparatus of the spinal cord. We say that the segmental apparatus has been pretuned, or simply, tuned, and that the detailed performance of coordinative structures is determined by the extant interactive state of the segmental apparatus. The tuning of coordinative structures and the activation of coordinative structures appear to be governed by separate mechanisms.

We now ask: If it is the case that the activation and tuning of coordinative structures are separately controlled events, at what level is the separation first evident? More precisely, we are keenly interested in the issue of whether tuning is the responsibility of the executive, and thus part of the initial representation of the act, or whether this responsibility lies outside the executive's domain.

For a given movement, such as plantar flexion, we may suppose that the executive specifies a tuning function to the servomechanisms for the (possibly) separate α and γ instructions to follow. The independence of movement-related tuning would arise, on this account, because the tuning function is effected by substructures different from those responsible for motor-neuron activation, much along the lines that the delimiting of coordinative structures and their decomposition are controlled separately. In this view, the family of possible tunings defines just another equivalence class, another invariant unit of information for the executive specification of solutions to action problems.

Alternatively, we may propose that segmental tuning is not specified in the action plan but is determined by other structures on acknowledgment of the executive's intention (cf. Greene, 1971a, 1971b). There would be special advantages accruing to a devolution of responsibility for specifying action plans and segmental tuning, advantages that would be especially pronounced when actions are related to environmental events. For example, it would mean that the executive could develop a repertoire of plans appropriate to frequently occurring classes of environmental events, so that when confronted with an event of a certain class, the executive issues a standard set of instructions and leaves to relatively independent tuning systems the responsibility for achieving the appropriate variant. Indeed, the largely invariant species-specific behavior of animals documented in the now celebrated works of ethologists (e.g., Tinbergen, 1951) strongly suggests that evolution has thoroughly exploited the principle of separating action-plan specification from tuning. The instinctive rituals are released by stimulation of a simple kind--the red belly of the stickleback, the spot under the herring-gull's beak--but the unfolding stereotypic behavior is flexible: it relates to the lay of the land, to the contingencies of the local environment. We should suppose that these species-specific action plans are adjusted by the pickup of information about the environment, that is to say, their tuning is environment related.

VISUAL CONTROL OF LOCOMOTION

Locomotion provides an instructive example of this point of view, for although locomotion propels an animal through its cluttered, textured environment, the basic locomotion pattern-generator is independent of local conditions (Evarts et al., 1971). The necessary adaptive modifications are effected by feedback from the peripheral motor apparatus (the muscles and the joints), from changes in tactual motion, from the basic orienting system (Gibson, 1966b), and most significantly, from the perceptual pickup of information about surfaces and objects, about the relations among them and the moving animal. Visually detected information about the environment plays a fundamental role in permitting anticipatory changes in the basic locomotion pattern through "feedforward"; appropriate changes in coordination may be induced before the animal confronts a certain kind of surface irregularity or a certain kind of object. To manipulate the locomotion plan by touch or kinesthetic feedback alone would be unsatisfactory, since this form of regulation would often occur after an ill-adjusted movement and thus would specify compensatory changes for states that are no longer current. It is far better to have the low-level realization of the plan adjusted beforehand through patterns of feedforward related to properties of the optic array and to leave to touch and joint-muscle feedback the task of achieving small, final adjustments. At all events, the locomotion illustration here raises the important issue of how the visual detection of environmental properties relating to the modification and control of locomotion is realized in the language of the action system.

With this issue in mind, let us proceed to examine in some detail the problem of how an animal moves about in a stable environment. We take as our orientation Gibson's (1958) analysis of locomotion and its control by vision. First we recognize, following Gibson, two fundamental assertions: the control of locomotion relative to the total environment is governed by transformations of the total optic array to a moving point; the control of locomotion relative to an object in the environment is governed by transformations of a smaller bounded cone of the optic array--a closed contour with internal texture in the animal's visual field. Second, and again respecting Gibson, we recognize the following as aspects of locomotion requiring our attention: beginning locomotion in a forward direction; terminating locomotion; locomoting in reverse; steering toward a specific location or object; approaching without collision; avoiding obstacles; pursuing and avoiding a moving object. Additionally, we recognize that locomotion must be adjusted to the physical properties of the surface--its convexities and concavities, its slants and slopes, its edges.

For each of the aspects of locomotion, we can identify correspondences in the flow patterns of the optic array. Thus to initiate locomotion in a forward direction is to activate and relate the coordinative structures that comprise the locomotor synergism (Gelfand et al., 1971) in such a fashion as to make the forward optic array flow outward; to cease locomotion is to terminate the optic flow; and to locomote in reverse is to pattern the locomotor synergism in a manner that makes the optic array flow inward. To move faster or slower is to make the rate of flow increase and decrease, respectively. As Gibson (1958:187) remarks: "An animal who is behaving in these ways is optically stimulated in the corresponding ways, or, equally, an animal who so acts to obtain these kinds of optical stimulation is behaving in the corresponding ways." Now, during forward movement, the center of the flow pattern is the direction in which the animal is moving, that is to say, the part of the array from which the optic flow pattern radiates corresponds to that part of the solid environment to which the animal is locomoting. If the animal changes direction, then naturally the center of flow shifts across the array. Thus we can say that to maintain locomotion in the direction of an object is to keep the center of flow of the optic array as near as possible to that part of the structure of the optic array that the object projects.

In moving about a stable environment, an animal will approach solid surfaces that it will need to contact or avoid as situation and history demand. Objects are specified in the optic array by contours with internal texture. Areas between objects are specified either by untextured homogeneous regions (e.g., sky) or by densely textured regions (e.g., sand, grass). In approaching an object, the closed contour in the array corresponding to the boundaries of the object expands with the rate of expansion for a uniform approach speed, accelerating in inverse proportion to the animal's proximity to the object. If the animal is on a collision course with the object, then a symmetrically expanding radial flow field will be kinetically defined over the texture bounded by the object's contours. On the other hand, if the expansion is skewed, i.e., if the pattern of texture flow is asymmetrical, then this specifies to the animal that it is on a noncollision course. A translation of the center of the flow pattern laterally to the animal's right or to the animal's left specifies that the animal will bypass the object on, respectively, its right or left. In Gibson's (1958) account, the guiding principle for approaching an object without collision is to move so as to cancel the forward and relatively symmetrically expanding flow of the optic array corresponding to the object at the instant when "the contour of the object on the texture of the surface reaches that angular magnification at which contact is made" (p. 188).[8] And to avoid objects, to steer successfully around them, the animal needs to keep the center of the centrifugal flow of the optic array outside the contours with internal texture and inside the homogeneous or densely textured surface areas.

Suppose now that the object to which movement is being directed is a moving object, as in the case of one animal pursuing another. We can again identify corresponding properties of the optic array. A prey fleeing a predator is specified by the fact that for the predator the overall optic array flows from a center, but a contour with internal texture within the overall flow pattern is not expanding: absolute expansion of the contour means that our predator is making good ground on his prey, contraction of the contour may mean that our prey will live to run another day. The principle of pursuit is summed up lightheartedly by Gibson (1958:188), "...the rule by which a big fish can catch a small fish is simple: maximize its optical size in the field of view."

We see, in short, that controlling locomotion calls for the detection of change, detection of rates of change, and detection of rates of rates of change in the flowing optic array. It also calls for the detection of changes in parts of the structure of the optic array with respect to the optic array as a whole.

---

[8] For an interesting experiment in insect behavior that is of some relevance to these comments and to the general theory of perception-action relations, see Goggshall (1972).

We assume that animals are sensitive to all of these properties of stimulation that vary over time and that they do indeed detect them (Gibson, 1966b; Ingle, 1968). We should also note that modulating the optic array through movement and modulating movement through changes in the optic array go hand in hand; thus the cybernetic loop of afference, efference, reafference is virtually continuous.

But we must now face up to a point that has been neglected thus far. In directing its locomotion to one object and weaving its way among others, and in pursuing one moving object and fleeing another, the animal exhibits its capacity to make discriminative responses. But these responses must be based on different properties of stimulation from those that determine the control of locomotion: they are responses specific to those properties of the optic array that do not change, as opposed to those that do; importantly, they are properties of stimulation that do not result from the animal's locomotion. The animal must be able to detect permanent properties of his environment: he must be able to detect whether a surface affords locomotion and whether a contour with internal texture affords collision; he must be able to detect whether a moving textured contour affords eating or whether it affords being eaten.

In respect to the surface supporting locomotion, the terrestrial animal must detect the gradients of optical texture specifying slant and slope, the topological shearing of texture specifying edge, and the changes of texture gradient specifying convexities and concavities. As he moves rapidly across a rough terrain, he must adjust his footfall pattern, temporally and spatially; he must adjust his gait to the wrinkled surface. He must detect surface protuberances and surface breaks requiring leaping-over, as opposed to those requiring going-round or avoiding; he will often need to make transitions between running and leaping. With respect to the permanent properties of the environment, we concur with Gibson (1966b) that the animal can detect in the changing optical flux those mathematically invariant properties that correspond to the physically constant object or surface and that afford for the organism possibilities for action.

We are led, therefore, to a distinction between those properties of stimulation that afford approach, avoidance, pursuit, flight, changes in the footfall pattern of a gait, and transitions from running to leaping, from those properties of stimulation that control locomotion in each of these respects. It would seem that the former are those properties that do not vary over time, while the latter are those properties that do. And the pickup of change and nonchange are concurrent perceptual activities.

TUNING REFLEXES AND ENVIRONMENT-RELATED TUNING

At this stage of our inquiry as to how vision enters into locomotion (and into action in general), we turn our attention to the concept of tuning reflexes. In addition to those reflexes that resemble parts of acts, such as the flexion and crossed-extension reflexes, or are themselves simple yet self-sufficient acts, such as the righting reflex and the scratch reflex, we can identify a further class of reflexes whose task, apparently, is to impose biases on the action system. We can distinguish therefore between "elemental" reflexes and "tuning" reflexes (Greene, 1969). As illustrations of tuning reflexes, we can take classically defined postural or attitudinal (Magnus, 1925) reflexes, such as the tonic neck reflex, which biases the motor apparatus for movement in the direction of gaze, and the labyrinthine reflexes, which bias the musculature to resist motion on an incline or to resist rotation (Roberts, 1967). Quite recently, evidence has been forwarded of low-level tuning resulting from movements of the eyes (Easton, 1971, 1972b). In the cat, stretching the horizontal eye muscles facilitates a turning of the neck and head from the direction of gaze, and stretching of the vertical eye muscles influences the forelimbs. Indeed, it appears that the eyes looking upward might foster forelimb flexion, and the eyes looking downward might foster forelimb extension (Easton, 1972a).

The principal function of tuning reflexes seems to be that of altering the intrinsic system of segmental relations rather than that of initiating configurations of motions in components of the motor machinery.[9] The impression is that tuning reflexes adjust the bias in the fundamental servomechanisms (cf. Gernandt, 1967). In general, it may be argued that the main advantage of tuning reflexes, whether induced by prior motion or induced more directly, is a reduction in the detail required of high-level instructions (Easton, 1972a). Thus, when a cat turns its head to gaze at a passing mouse, the angle of tilt of the head and the degree of flexion and torsion in the neck will elicit a reflex modulation of the segmental apparatus such that a broadly stated executive instruction to "jump" will be realized as a jump in the right direction (Magnus, 1925; Ruch, 1965b). Clearly, such modulation must precede the innervation of muscles or the cat would constantly miss its target; obviously, the cat in flight cannot rely on corrective feedback.

How do tuning reflexes relate to the visual control of locomotion? Analysis of the biomechanics of walking and running in animals (e.g., Arshavskii, Kots, Orlovskii, Rodionov, and Shik, 1965; Shik and Orlovskii, 1965; Shik, Orlovskii, and Severin, 1966) reveals that with change in speed or gait the majority of kinematic parameters is kept constant, suggesting that adjustments in the locomotion plan require a relatively minimal change in coordination. The action problem posed by the need to change speed of running or gait may be solved in most instances by a change in only two parameters. May we suppose therefore that a change in a small set of parameters is all that is needed to control locomotion through a "wrinkled" and object-cluttered terrain? Movement in a forward direction calls for a particular organization of the basic coordinative structures. If an animal so moving detects an invariant specifying an object or surface in its path that is to be avoided, then it must alter the organization of the relevant coordinative structures in order to change direction. But change

---

[9] The potential range of changes in the segmental system induced by postural changes, and their implications for the behavior of coordinative structures, is suggested in the following paragraph from an address delivered by Magnus (1925: 346) fifty years ago: "Every change in attitude, with its different positions of all parts of the body, changes the reflex excitability of these parts and in some cases changes also the sense of the reflex evoked, excitations being converted into inhibitions, reflex extensions into flexions and so on. One and the same stimulus applied to one and the same place on the body may give rise to very different reactions in consequence of different attitudes which have been imposed to the body before the stimulus is applied." For further intriguing and provocative comments on tuning reflexes, see Jones (1965) and Fukuda (1957).

in direction need not actually require direct executive intervention in the low-level organization of the locomotion plan; a shift in the direction of gaze may be all that is needed. In theory at least the tonic neck reflexes and related tuning reflexes could effect the necessary reorganization of the segmental apparatus. Similarly, if the contoured texture in the optic array afforded jumping-on then the act of directing the eyes, or eyes and head, upward would facilitate the transition in segmental organization from that of running to that of jumping.

These examples suggest the following: in the course of locomotion, the detection of invariants affording specific changes in locomotion may serve to activate singularly simple action plans such as a change in the direction in which the head and/or eyes are pointing. Often these adjustments in orientation--owing to the functional tuning link between head and eye movements and the segmental apparatus--are sufficient to produce the needed parameters for the segmental realization of change in locomotion.

As we have noted, the optical stimulation for a moving animal has components of both change and nonchange (Gibson, 1966b). If the components of nonchange, specifying the permanent entities in the animal's environment, relate to action plans and their activation, to what do the changing components of stimulation relate? We must suppose that they relate to the mechanisms of tuning; but how is this relation effected? The following considerations may help us to move toward an answer to this question.

To leap from object to object is to project the body in particular trajectories, with each trajectory requiring different horizontal and vertical vectors of extension thrust. Variations in force could be achieved either through variations in the degree of activation of coordinative structures or parts of coordinative structures as might be permitted by the local sign properties of reflexes (the dependency of reflex patterns on the origin of stimulation) or through direct facilitation of motor-neuron activity, or both (cf. Easton, 1972a). In theory, both of these sources of force variation are plausible instances of tuning. Therefore, we can say that each leap calls for the specification of parameters to the intrinsic system of segmental relations where these parameters relate to the desired trajectory. Now we might ask whether trajectory-related parameters could be determined through tuning reflexes. But cursory analysis would suggest that mechanical modulation--spinal tuning elicited reflexively by a prior motion such as directing the eye-head system toward an object--is inadequate for the task. Consider a cat perched on a particular platform. At a distance of X feet from his perch is another, higher platform. Directing his gaze to the top surface of the higher platform yields, say, a particular angle of neck extension and hence a particular tuning of the segmental apparatus. Yet we observe that we could arrange any number of higher platforms of different heights at any number of reasonable distances either more or less than X feet from the cat's perch that would correspond to the same inclination of the neck and hence to the same tuning parameters and hence, supposedly, to the same degree of thrust if the cat chose to jump. In brief, reflex tuning induced by any particular orientation of the eye-head system is ambiguous with respect to distance. Mechanically induced tuning, therefore, cannot supply the tuning parameters relevant to a given trajectory. How then are they supplied? We are forced to conclude that they are supplied by the properties of the optic array that specify relative distance and height in the cat's normal cluttered and textured environment, and that these optical properties are realizable as segmental tunings without the intervention of executive procedures and without mechanical mediation.

With this conclusion in mind, consider what we might now say about the scenario that unfolds when a scampering mouse appears at a leapable distance from an interested cat. In the cat's field of vision, the mouse is projected as changing patterns in the optic array. Concurrently, there is a pickup by the cat's visual system of those properties of stimulation that change over time and those properties that do not. The former specify how far away the mouse is, in what direction it is moving, at what rate it is moving, and where it will be in a following instant relative to the cat; the latter specify the mouse's identity as something that affords catching and eating. Orienting in the direction of the mouse adjusts the segmental apparatus through the tuning reflexes for a movement in that direction; as the direction of gaze shifts according to changes in the mouse's location, the mechanically induced segmental tuning likewise adjusts apropos the new direction. On activation of the action plan to pounce, the tuning parameters for the needed trajectory specified by the transformations in the optic array are given to the segmental system of interactions. The activation of coordinative structures then takes place against a backdrop of segmental relations appropriately adjusted for the generation of a precise, on-target leap.

In this cat-and-mouse story there are two main themes: one is that the activation of crudely stated action plans and environment-related tuning are based on different properties of stimulation; the other is that the properties of visual stimulation that control movement and the family of possible tunings that effect the control are tightly linked. In Gibson's view, perception is direct. He has also remarked that: "The distinction between an S-R theory of control reactions and an S-R theory of identifying reactions is important for behavior theory" (Gibson, 1958:190). On this distinction, we might now comment that in control reactions the relation is between changing properties of stimulation and patterns of tuning, and in identifying reactions it is between nonchanging properties of stimulation and action plans.

Mittelstaedt (1957) describes a similar story about prey capture in the mantis. A mantis strikes its prey with pinpoint accuracy within a latency of 10 to 30 msec, a period too brief to allow for adjustments during the course of the strike trajectory. The problem is to account for how this accuracy is achieved when the prey appears at a strikable distance either to the left or to the right of the body axis at some variable angle; and when the head is oriented at some (different) angle to the prothorax with which the forelegs--the striking instrument--are articulated. Mittelstaedt's modeling of this situation implies that the visual and proprioceptive information specifying the relevant relations is conveyed not to the executive issuing the strike signal but to the segmental machinery of the forelegs. On our account, we would say that the higher-order invariant specifying "prey" triggers the strike command (the strike action plan), but the properties of optical stimulation specifying the coordinates of the prey with respect to the body axis, and its rate and direction of movement, do not enter into the executive decision, for most assuredly that would introduce undesirable delays. Rather, these properties are realized as segmental tuning parameters effecting needed adjustments in the centers controlling foreleg extension. We may say of the mantis' prey-catching that the prey determines the ballistic component, while the prey's location and movement determine the tuning component in a mixed ballistic-tuning strategy. Moreover, we recognize what might indeed be a general principle, namely, that different properties of stimulation enter into the unfolding act at different levels.

In respect to this last point, let us make one final comment on the topic of locomotion, which began this particular phase of our inquiry. We have argued that environment-related tuning is relatively independent of executive procedures. For locomotion, we can say that tuning is coupled to the pickup of information conveyed by continuous transformations in the optic array. While the detection of higher-order invariances (affordances) may inject gross adjustments in locomotor activity, the fine control of locomotion in an object-cluttered and wrinkled terrain is through environment-related tuning, which adapts the activity to the conditions by modulating a relatively small set of parameters, and does so without involving the higher domains of the action coalition/heterarchy.

Pal'tsev (1967a, 1967b) advanced a theory of special relevance to this account of locomotor regulation. First, Pal'tsev (1967a) recognizes that, in respect to uniform movements, an argument can be made that, in addition to movement-related segmental pretuning, there is another type of reorganization of the segmental relations that is brought about during the execution of the movements. In Pal'tsev's (1967a) view, this latter form of tuning is largely due to the fact that the interactions among different structures of the spinal cord are reorganized by processes that are inherently spinal. The segmental apparatus tunes itself, as it were, in harmony with the main supraspinal influence. By comparing experimental results on the effects spinal reflexes induce in neighboring spinal reflexes with the general picture of locomotion, Pal'tsev (1967b) is led to the supposition that, following the first few locomotor cycles, the strategic ordering of muscle events in locomotion can be determined solely by the segmental system of relations. As he sees it, the supraspinal patterns of feedforward serve only to identify and to "trigger" the particular locomotion plan; the continuation of the plan--the subsequent locomotor cycles in walking or running--is then the responsibility of spinal processes. That is to say that control of locomotion is simply and elegantly transferred from supraspinal structures to spinal structures. Thus, locomotion exhibited in the pursuit by a predator of its prey could proceed with insignificant involvement of the highest sectors of the action system. If such is the case, then it would be propitious for the nervous system to exploit the principle of conveying visually specified adjustments in locomotion relatively directly to the segmental apparatus in which locomotion control is invested. This conclusion is consonant with the point of view often expressed by Russian investigators that the spinal cord is a system that during action serves to integrate different supraspinal influences (cf. Pal'tsev and El'ner, 1967).

TWO KINDS OF VISION

In some respects the ideas just expressed are reminiscent of the claim that there are two separate but interdependent visual systems related to action (Trevarthan, 1968). It appears that a distinction is drawn in the neuroanatomy of the brain "between vision of relationships in an extensive space and visual identification of things" (Trevarthan, 1968:301). In its simplest form, the distinction is demonstrated most straightforwardly by the experiments of Schneider (1969): a hamster with intact superior colliculus but no visual cortex can orient to objects but cannot distinguish between them; conversely, with intact visual cortex and no superior colliculus the hamster can successfully distinguish objects but cannot locate them and orient to them except through trial-and-error. In very general terms, it appears that there is a functional differentiation between two kinds of vision that relates in part to forebrain-midbrain differences.

Let us remark briefly on the vertebrate midbrain. Suppose that we drew a map of the projections from the eyes to the midbrain tectum and suppose that we did so for two dissimilar vertebrates, the goldfish and the cat. The eyes of the goldfish are aligned roughly perpendicular to the body axis, while those of the cat are aligned parallel to the body axis. If we drew our maps in the optical coordinates of the eye, we would find that for our two vertebrates the projection from the eyes to the midbrain differed considerably. But if our maps were drawn in the coordinates of the behavioral field, that is, with respect to the symmetry of the body, we would observe that the two maps were virtually identical. Indeed, if we went on to obtain such maps for other vertebrates, we would find that the mapping from eyes to tectum in the coordinates of the behavioral field is relatively invariant, and thus indifferent to the variation in alignment between eyes and body axis (see Trevarthan, 1968). One might conjecture that body-centered visual space is represented by a precise topographical mapping in the midbrain in very much the same way in all vertebrates.

This map of visual loci also maps a topography of points of entry into the action system. Stimulating points on the tectum produces orienting movements of eyes, head, and trunk to the corresponding visual location (cf. Apter, 1946; Hyde and Eliasson, 1957; Ewert, 1974). A singularly important feature of the midbrain is that, in respect to the symmetry of the body, it provides a precise topographical map of points in visual space and a virtually identical map of orienting movements to those points (Apter, 1946). Because of this feature, the midbrain serves to map object locations onto the set of movement-induced tunings. But there is reason to suppose that the capabilities of the midbrain extend beyond this and are concerned in a more general way with the control of locomotion.

Let us say that the two kinds of vision relating to forebrain-midbrain differences relate in turn to different kinds of acts performed in the animal's behavioral space. Discussing primates, Trevarthan (1968:302) conveys the tenor of this point of view as follows: "Orientations of the head, postural adjustments, locomotor displacements change the relationship between the body and spatial configurations of contours, surfaces, events, and objects. These movements occur in what I shall call ambient vision. In contrast, praxic actions on the environment to use pieces of it in specific ways are performed with the motor apparatus of the body and the visual receptors oriented together so that both vision and the acts inflicted on the environment occur in one part of the behavioral space. The vision applied to one place and a specific kind of object, or deployed in a field of identified objects, I shall call focal vision."

Trevarthan builds his case on facts found in the effects of surgically separating the cerebral hemispheres. This separation exhibits many instructive and curious phenomena, including that of central concern to Trevarthan's thesis--the capability of the split-brain primate to double-perceive and learn for some types of visual stimulation but not for others and correspondingly to perform some aspects of visually defined acts chaotically and yet to perform others with no evident impairment. We note that the separated cortices may learn, independently and simultaneously, conflicting solutions to a visual discrimination problem when the stimuli are clearly of different identities, as in the example of cross versus circle, but not when the stimuli differ on a single dimension, such as bright versus dim. In the former case, that of an identity difference, what is learned by one hemisphere is available to the other only if it in turn has the opportunity to learn the same thing; in the latter case, what is learned by one hemisphere is without practice available to the other. The inference is that differences in degree may be apprehended by visual mechanisms of the midbrain, while the apprehension of differences in identity is the responsibility of cortical visual mechanisms (Trevarthan, 1968). And Trevarthan emphasizes that it is the transformations relating the to-be-distinguished stimuli, rather than the ease of distinguishing between them, that is important to the dissociation.

Paralleling this dissociation in split-brain vision is a dissociation in split-brain action. If an object such as a peanut is presented to the commissurectomized primate, both hands may reach forward with precision to grasp it; however, the activities of the two hands appear indifferent to each other, resulting often in collision. Given an object to manipulate and explore, the split-brain displays an inability to relate the activities of the two hands. The needed collaboration is replaced by redundant and conflicting movements.

In sharp contrast to these anomalies of voluntary movements of the hands in the field of focal vision, no such schism is witnessed in locomotion in which the hands play an important role. Locomotion-related movements of the arms and hands are properly coordinated to each other and to the motions of the hind limbs; and in terms of displacement, velocity and timing are finely attuned to the environmental structures supporting the action (Trevarthan, 1968).

While there are many more questions to be asked of these dissociations in vision and action and of the relation between them, we can with some reasonable certitude draw the following conclusions. First, low-level sections of the visual system can effect the pickup of transformations in the optic array corresponding to changes in gross environmental properties, such as texture and contour, and to the detection of simple invariances such as solidity--in short, those properties of the optic array relevant to the control of locomotion. And in this regard, it is of some import to note that electrical stimulation of the midbrain can bring about parameter changes in the segmental functions governing locomotion (Shik, Severin, and Orlovskii, 1966). Second, the higher-level sections of the visual system detect higher-order invariants specifying identity and more complex transformations that would be relevant to and indeed result from the skilled manipulation of objects. For it is evident that separating the hemispheres gives rise to two separate visual frames for the regulation of manipulative behavior and to a consequent breakdown in coordination between the two hands, but leaves intact the visual frame for the regulation of locomotion.

RELATING THE CONTENTS OF VISION TO ACTION: A SUMMING UP

We come now to a general summary of these speculations on how vision enters into action. We have provided two rather different descriptions of the unfolding act, and it will be helpful to collect them together at this time. In one, we envisaged the act as evolving through the establishment of progressively less abstract representations, from the specification of relations among and within a few relatively large coordinative structures to the specification of relations among and within many relatively small coordinative structures, the fundamental servomechanisms. In the other, we saw the action plan unfolding as ordered successions of progressively less abstract projections of the environment. What we must attempt now is a reconciliation of these separate views.

The kernel notion in this essay has been the idea of building acts through the fitting together of relatively autonomous units. This principle of operation reflects the fundamental argument that there are far too many degrees of freedom in coordinated activity for it to be controlled by a single procedure in a single instant. One consequence of this point of view is that the initial representation of an act in the highest domain must necessarily be crude in comparison to its ultimate representation in terms of instructions to muscles.

Similarly, we saw that in view of the degrees-of-freedom problem, the representation in the highest domain could not be constructed in respect to the details of skeletal space; the perception of the disposition of the limbs and branches of the body at any moment can only enter into the representation in the most general way as an abstracted account of the body's "pose" at that instant. Using very much the same rationale, we are led to believe that in interactions with the environment not all the contents of vision can be involved in the determination of the initial representation of an act. Again, we suppose that the executive procedure uses only the perceptual description that it can handle; the description cannot be detailed and by necessity must be fairly abstract. Earlier, following Bernstein (1967), we used the term "topological properties" to identify the description of the environment to which the initial representation of the action plan related. We may now regard these properties as invariants of a higher order: for example, those that specify the identities of objects and their possibilities of transformation. In any event, the manifestation of the action plan as motions finely attuned to the nuances of the environment's structure tells us quite plainly that the detailed contents of vision must be interjected into the act during its evolution. We have argued that tuning of coordinative structures is probably the mechanism through which the interjection of environmental details is brought about.

If these speculations are not too far off the mark, then we might further conjecture as follows. The determination of an act as an orderly pattern of motions is distributed across many structures. In the coalitional/heterarchical language used above, we say that it is distributed across different domains. But where the differentiation of an action plan requires information about the environment, we should suppose that the procedures operating at each domain incorporate optically specified environmental properties. It seems unlikely, however, that the entry of environmental properties into the various representations of an unfolding act is a haphazard affair. Rather, we hypothesize that the properties of the optic array interlace with the representation of an action plan in a systematic fashion: different properties map into different representations. We have, of course, already implied that this might be the case in arguing that the specification of action plans and the tuning of structures correspond to different properties of stimulation. But now suppose that the properties of stimulation relevant to the control of action may themselves be arranged from more complex to more simple; then perhaps we can imagine a natural mapping of these properties onto the unfolding act, a mapping that preserves their order. Of the properties of optical stimulation relevant to the control of action, those of a higher order are realized as tunings in higher domains and those of a lower order are realized as tunings in lower domains of the action coalition/heterarchy.

We conclude with some final thoughts on the general characterization of the perception-action relation. With respect to the representation of the action plan in the highest domain, it is not so much that the specification of a subset of large coordinative structures and functions defined on them relate to higher-order properties of the optic array but rather that the description of the plan in action terms and the description of the plan in perceptual terms are dual statements about the same thing. Earlier we described the action concept for A-writing as an operator defined over a set of functions relevant to the manipulation of coordinative structures; but we have also referred to the action concept for A-writing in geometrical terms consonant with the points of view expressed by Bernstein (1967) and Lashley (1951), and further suggested that the two descriptions were isomorphic. Similarly, with respect to tuning, we have implied that there is a relatively direct mapping of the properties of optical stimulation relevant to the control of action onto the set of tunings. To draw these concepts together, we can say that "detection of control-relevant optical properties" and "specification of environment-related tuning parameters" are descriptions of the same event: one is the dual of the other.

Perhaps we can gain a purchase on the duality of perception and action events by considering a problem drawn from a rather special domain of the perception-action relation--communication between members of the same species. For a variety of reasons, it has been suggested that the perception of the sounds of speech is achieved by reference to the mechanisms of articulation (see Galunov and Chistovich, 1966; Liberman et al., 1967; Zinkin, 1968). One version of this action-based theory of speech perception suggests that the listener seeks to determine (tacitly and unconsciously of course) which phoneme articulation plans could produce the acoustic pattern; the listener uses the inconstant sound to recover the articulatory gestures that produced it and thereby arrives at the speaker's intent (Liberman et al., 1967). Other students of speech, however, have argued against the articulatory matching explanation of the perception of speech sounds and have suggested that the explanation be sought in the sensitivity of the nervous system to higher-order properties of acoustic stimulation (e.g., Fant, 1967; Abbs and Sussman, 1971). There is growing evidence for neural mechanisms that selectively respond to complex acoustic invariances (e.g., Roeder and Treat, 1961; Frishkopf and Goldstein, 1963; Capranica, 1965) and it is becoming increasingly less venturesome to propose that the perception of phonological attributes of speech is direct rather than mediated (cf. Abbs and Sussman, 1971). However, viable descriptions of invariances in speech stimulation have been elusive.

Commendable as a direct perception interpretation is, we still must account for the evidently tight coupling between structures detecting speech sounds and structures producing speech sounds (see Chistovich, 1961; Chistovich, Fant, de Serpa-Leitaõ, and Tjernlund, 1966). Suppose, as Gibson (1966b) suggests, that vibratory patterns specify their source. Then we can say that a listener perceives articulation because the invariants of vibration correspond to the invariants of articulation: the phonemes are present in the neural activity and vocal-tract activity of the speaker and in the air between the speaker and the listener. Thus the linguistically relevant invariances on the input side are the same as the linguistically relevant variables on the output side, and it is in this sense that perceiving and producing speech correspond. Now suppose that we were to describe an articulatory action plan as a set of relations defined over a collection of coordinative structures, then we should argue that our description is also a description of the relevant relations in the acoustic pattern. An appropriate analogy is the group concept in mathematics: given two different sets of elements, with a group structure defined on each, we might find that although the elements differ (even radically) in the two instances, their manner of inner interlocking is the same, in which case we say that they represent the same abstract group. Our hypothesis, therefore, is this: the structure that affords perception of a speech sound also affords its production; speech perception and speech production are related by abstract structures that are common to both but indigenous to neither (cf. Turvey, 1974). There is some evidence, though slight, that structures with this dual property may have been exploited in the evolution of intraspecies communication. For example, the calling song of male crickets is composed of stereotyped rhythmic pulse intervals. Cross-breeding of two species of crickets with marked differences in the rhythmic structure of their songs produces hybrids whose calling song is distinctly different from either parental song. It has been shown that genetic differences that cause song change in males also alter song reception in the females: hybrid females prefer the song of hybrid males (Hoy and Paul, 1973). Especially intriguing is the speculation that the action plan for song generation in the male and the female's selective sensitivity to the male's song are coupled through a common set of genes (Hoy and Paul, 1973). Thus, at some level of abstraction, the same structure may underlie song production in the male and song reception in the female.

Whether a stronger and more general case can be made for the dual representation notion remains to be seen. There is the possibility, of course, that the principle we have tried to describe has meaning, if at all, only in the communication mode: speaking and perceiving speech, reading and writing, and the primitive instantiations of signaling in animal and insect communication. On the other hand, when one considers the failure of schemes in which sensory input is routed through a central network into motor responses, the growing uneasiness over the application of the terms "sensory, motor, associative" to higher neural structures, the increasing usage of the bimodal term "sensorimotor" (see Evarts et al., 1971, for comments on each of these points), and the arbitrariness of action-based theories of perception, then the notion of perceiving and acting as dual representations of common neural events may be a reasonable alternative to the sensory and motor views of mind.

REFERENCES

Abbs, J. H. and H. M. Sussman. (1971) Neurophysiological feature detectors and speech perception: A discussion of theoretical implications. J. Speech Hearing Res. 14, 23-36.

Apter, J. T. (1946) Eye movements following strychninization of the superior colliculus of cats. J. Neurophysiol. 9, 73-86.

Arbib, M. A. (1972) The Metaphorical Brain: An Introduction to Cybernetics as Artificial Intelligence and Brain Theory. (New York: Wiley).

Arshavskii, Yu. I., Ya. M. Kots, G. N. Orlovskii, I. M. Rodionov, and M. L. Shik. (1965) Investigation of the biomechanics of running by the dog. Biophys. 10, 737-746.

Asatryan, D. G. and A. G. Fel'dman. (1965) Functional tuning of the nervous system with control of movement or maintenance of a steady posture - 1. Mechanographic analysis of the work on the joint on execution of a postural task. Biophys. 10, 925-935.

Bartlett, F. C. (1964) Remembering. (Cambridge: Cambridge University Press).

Belen'kii, V. Yi., V. S. Gurfinkel, and Ye. I. Pal'tsev. (1967) Elements of control of voluntary movements. Biophys. 12, 154-161.

Bernstein, N. (1967) The Co-ordination and Regulation of Movements. (London: Pergamon Press).

Capranica, R. R. (1965) The evoked vocal response of the bullfrog: A study of communication by sound. Research Monographs, No. 33 (Cambridge, Mass.: MIT Press).

Care, N. S. and C. Landesman. (1968) Readings in the Theory of Action. (Scarborough, Ont.: Fitzhenry and Whiteside).

Cassirer, E. (1957) The Philosophy of Symbolic Forms, Vol. 3 of The Phenomenology of Knowledge. (New Haven: Yale University Press).

Chistovich, L. A. (1961) Classification of rapidly repeated speech sounds. Sov. Phys. Acoust. 6, 393-398.

Chistovich, L., G. Fant, A. de Serpa-Leitaõ, and P. Tjernlund. (1966) Mimicking of synthetic vowels. Quarterly Progress Status Report (Speech Technology Laboratory, Royal Institute of Technology, Stockholm, Sweden) QPSR 2/1966.

Chomsky, N. (1965) Aspects of the Theory of Syntax. (Cambridge, Mass.: MIT Press).

Chomsky, N. (1966) Topics in the Theory of Generative Grammar. (The Hague: Mouton).

Easton, T. A. (1971) Patterned inhibition from horizontal eye movement in the cat. Exp. Neurol. 31, 419-430.

Easton, T. A. (1972a) On the normal use of reflexes. Amer. Sci. 60, 591-599.

Easton, T. A. (1972b) Patterned inhibition from single eye muscle stretch in the cat. Exp. Neurol. 34, 497-510.

Eccles, J. C. (1969) The dynamic loop hypothesis of motor control. In Information Processing in the Nervous System, ed. by K. N. Leibovic. (New York: Springer-Verlag).

Eccles, R. M. and A. Lundberg. (1959) Supraspinal control of interneurones mediating spinal reflexes. J. Physiol. 147, 565-584.

Eldred, E. (1960) Posture and locomotion. In Handbook of Physiology: Neurophysiology, Vol. II, ed. by H. W. Magoun. (Washington, D.C.: American Physiological Society).

Engberg, I. and A. Lundberg. (1969) An electromyographic analysis of muscular activity in the hindlimb of the cat during unrestrained locomotion. Acta Physiol. Scand. 75, 614-630.

Evarts, E. V. (1967) Representation of movements and muscles by pyramidal tract neurons of the precentral motor cortex. In Neurophysiological Basis of Normal and Abnormal Motor Activities, ed. by M. D. Yahr and D. P. Purpura. (New York: Raven Press).

Evarts, E. V. (1973) Brain mechanisms in movement. Sci. Amer. 229(1), 96-103.

Evarts, E. V., E. Bizzi, E. E. Burke, M. Delong, and W. T. Thach. (1971) Central control of movement. Neurosci. Res. Prog. Bull. 9, No. 3.

Evarts, E. V. and W. T. Thach. (1969) Motor mechanisms of the CNS: Cerebrocerebellar interrelations. Ann. Rev. Physiol. 31, 451-489.

Ewert, J-P. (1974) The neural basis of visually guided behavior. Sci. Amer. 230(3), 34-42.

Falk, G. (1972) Interpretation of imperfect line data as a 3-dimensional scene. Artifi. Intell. 3, 101-144.

Fant, G. (1967) Auditory patterns of speech. In *Models for the Perception of Speech and Visual Form*, ed. by W. Wathen-Dunn. (Cambridge, Mass.: MIT Press).

Fel'dman, A. G. (1966a) Functional tuning of the nervous system with control of movement or maintenance of a steady posture - II. Controllable parameters of the muscles. Biophys. 11, 565-578.

Fel'dman, A. G. (1966b) Functional tuning of the nervous system with control of movement or maintenance of a steady posture - III. Mechanographic analysis of the execution by man of the simplest motor tasks. Biophys. 11, 766-775.

Festinger, L., C. A. Burnham, H. Ono, and D. Bamber. (1967) Efference and the conscious experience of perception. J. Exp. Psychol. Monogr. 74(4, Pt. 2).

Frishkopf, L. and M. Goldstein. (1963) Responses to acoustic stimuli from single units in the eighth nerve of the bullfrog. J. Acoust. Soc. Amer. 35, 1219-1228.

Fukuda, T. (1957) *Stato-kinetic Reflexes in Equilibrium and Movement*. (Tokyo: Igaku Shoin).

Galunov, V. I. and L. A. Chistovich. (1966) Relationship of motor theory to the general problem of speech recognition (review). Sov. Phys. Acoust. 11, 357-365.

Gelfand, I. M., V. S. Gurfinkel, M. L. Tsetlin, and M. L. Shik. (1971) Some problems in the analysis of movements. In *Models of the Structural-Functional Organization of Certain Biological Systems*, ed. by I. M. Gelfand, V. S. Gurfinkel, S. V. Fomin, and M. L. Tsetlin. (Cambridge, Mass.: MIT Press).

Gernandt, B. E. (1967) Vestibular influence upon spinal reflex activity. In *Myotatic, Kinesthetic, and Vestibular Mechanisms*, Ciba Foundation Symposium. (London: Churchill).

Gibson, J. J. (1958) Visually controlled locomotion and visual orientation in animals. Brit. J. Psychol. 49, 182-194.

Gibson, J. J. (1966a) The problem of temporal order in stimulation and perception. J. Psychol. 62, 141-149.

Gibson, J. J. (1966b) *The Senses Considered as Perceptual Systems*. (Boston: Houghton Mifflin).

Goggshall, J. C. (1972) The landing response and visual processing in the milkweed bug, *Oncopeltus fasciatus*. J. Exp. Biol. 57, 401-414.

Gottlieb, G. L., G. C. Agarwal, and L. Stark. (1970) Interaction between voluntary and postural mechanisms of the human motor system. J. Neurophysiol. 33, 365-381.

Greene, P. H. (1969) Seeking mathematical models for skilled actions. In *Biomechanics*, ed. by D. Bootzin and H. C. Muffley. (New York: Plenum Press).

Greene, P. H. (1971a) Introduction. In *Models of the Structural-Functional Organization of Certain Biological Systems*, ed. by I. M. Gelfand, V. S. Gurfinkel, S. V. Fomin, and M. L. Tsetlin. (Cambridge, Mass.: MIT Press).

Greene, P. H. (1971b) Problems of organization of motor systems. Quarterly Report No. 29 (Institute for Computer Research, University of Chicago).

Gunkel, M. (1962) Über relative koordination bei willkürlichen menschlichen Gliedbewegungen. Pflügers Archiv Für Die Gesamte Physiologia 215, 472-477.

Gurfinkel, V. S., Ya. M. Kots, V. I. Krinskiy, Ye. I. Pal'tsev, A. G. Fel'dman, M. L. Tsetlin, and M. L. Shik. (1971) Concerning tuning before movement. In *Models of the Structural-Functional Organization of Certain Biological Systems*, ed. by I. M. Gelfand, V. S. Gurfinkel, S. V. Fomin, and M. L. Tsetlin. (Cambridge, Mass.: MIT Press).

Gurfinkel, V. S. and Ye. I. Pal'tsev. (1965) Effect of the state of the segmental apparatus of the spinal cord on the execution of a simple motor reaction. Biophys. 10, 944-951.

Haber, R. N. (1969) Information processing analyses of visual perception: An introduction. In Information Processing Approaches to Visual Perception, ed. by R. N. Haber. (New York: Holt, Rinehart & Winston).

Hayek, F. A. (1969) The primacy of the abstract. In Beyond Reductionism, The Alpbach Symposium, ed. by A. Koestler and J. R. Smythies. (Boston: Beacon Press).

Hoffman, P. (1922) Untersuchungen über die Eigenreflexe (Sehnreflexe) Menschlicher Muskeln. (Berlin: Springer).

Hoy, R. R. and R. C. Paul. (1973) Genetic control of song specificity in crickets. Science 180, 82-83.

Hyde, J. E. and S. G. Eliasson. (1957) Brainstem induced eye movements in cats. J. Comp. Neurol. 108, 139-172.

Ingle, D. (1968) Spatial dimensions of vision in fish. In The Central Nervous System and Fish Behavior, ed. by D. Ingle. (Chicago: University of Chicago Press).

James, W. (1890) The Principles of Psychology. (New York: Holt).

Jones, F. P. (1965) Method for changing stereotyped response patterns by the inhibition of certain postural sets. Psychol. Rev. 72, 196-214.

Koestler, A. (1969) Beyond atomism and holism--the concept of the holon. In Beyond Reductionism, The Alpbach Symposium, ed. by A. Koestler and J. R. Smythies. (Boston: Beacon Press).

Kornhuber, H. H. (1974) Cerebral cortex, cerebellum, and basal ganglia: An introduction to their motor functions. In The Neurosciences: Third Study Program, ed. by F. O. Schmitt and F. G. Worden. (Cambridge, Mass.: MIT Press).

Kots, Ya. M. (1969a) Supraspinal control of the segmental centres of muscle antagonists in man - I. Reflex excitability of the motor neurones of muscle antagonists in the period of organization of voluntary movement. Biophys. 14, 176-183.

Kots, Ya. M. (1969b) Supraspinal control of the segmental centres of muscle antagonists in man - II. Reflex excitability of the motor neurones of muscle antagonists on organization of segmental activity. Biophys. 14, 1146-1154.

Kots, Ya. M., V. I. Krinskiy, V. L. Naydin, and M. L. Shik. (1971) The control of movements of the joints and kinesthetic afferentation. In Models of the Structural-Functional Organization of Certain Biological Systems, ed. by I. M. Gelfand, V. S. Gurfinkel, S. V. Fomin, and M. L. Tsetlin. (Cambridge, Mass.: MIT Press).

Kots, Ya. M. and A. V. Syrovegin. (1966) Fixed set of variants of interactions of the muscles of two joints in the execution of simple voluntary movements. Biophys. 11, 1212-1219.

Kots, Ya. M. and V. I. Zhukov. (1971) Supraspinal control of the segmental centres of muscle antagonists in man - III. "Tuning" of the spinal apparatus of reciprocal inhibition in the period of organization of voluntary movement. Biophys. 16, 1129-1136.

Kuno, M. and E. R. Perl. (1960) Alteration of spinal reflexes by interaction with suprasegmental and dorsal root activity. J. Physiol. 151, 103-123.

Lashley, K. S. (1951) The problem of serial order in behavior. In Cerebral Mechanisms in Behavior, The Hixon Symposium, ed. by L. A. Jeffress. (New York: Wiley).

Lenneberg, E. (1967) Biological Foundations of Language. (New York: Wiley).

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Lundberg, A. (1969) Reflex control of stepping. In Proceedings of Norwegian Academy of Science and Letters. (Oslo: Universitetsforlaget).

Magnus, R. (1925) Animal posture. Proc. Roy. Soc. London 98(Ser. B), 339-353.

Matthews, P. B. C. (1964) Muscle spindles and their motor control. Physiol. Rev. 44, 219-288.

Merton, P. A. (1972) How we control the contraction of our muscles. Sci. Amer. 226(5), 30-37.

Miller, G. A., E. Galanter, and K. H. Pribram. (1960) Plans and the Structure of Behavior. (New York: Henry Holt).

Minsky, M. and S. Papert. (1972) Artificial intelligence. Artificial Intelligence Memo (Artificial Intelligence Laboratory, MIT, Cambridge, Mass.) 252.

Mittelstaedt, H. (1957) Prey capture in mantids. In Recent Advances in Invertebrate Physiology, ed. by B. T. Scheer. (Eugene, Ore.: University of Oregon Press).

Navas, F. and L. Stark. (1968) Sampling or intermittency in hand control system dynamics. Biophys. J. 8, 252-302.

Neisser, U. (1967) Cognitive Psychology. (New York: Appleton-Century-Crofts).

Obituary: M. L. Tsetlin. (1966) Biophys. 11, 1080.

Oscarsson, O. (1970) Functional organization of spinocerebellar paths. In Handbook of Sensory Physiology, Vol. II--Somato-sensory System, ed. by A. Iggo. (Berlin: Springer-Verlag).

Paillard, J. (1960) The patterning of skilled movements. In Handbook of Physiology: Neurophysiology, Vol. 3, ed. by J. Field, H. W. Magoun, and V. E. Hall. (Washington, D.C.: American Physiological Society).

Pal'tsev, Ye. I. (1967a) Functional reorganization of the interaction of the spinal structure in connexion with the execution of voluntary movement. Biophys. 12, 313-322.

Pal'tsev, Ye. I. (1967b) Interactions of the tendon reflex areas in the lower limbs in man as a reflexion of locomotor synergism. Biophys. 12, 1048-1059.

Pal'tsev, Ye. I. and A. M. El'ner. (1967) Change in the functional state of the segmental apparatus of the spinal cord under the influence of sound stimuli and its role in voluntary movement. Biophys. 12, 1219-1226.

Patton, H. D. (1965) Reflex regulation of movement and posture. In Physiology and Biophysics, ed. by T. C. Ruch and H. D. Patton. (Philadelphia: W. B. Saunders).

Pribram, K. H. (1971) Languages of the Brain: Experimental Paradoxes and Principles in Neuropsychology. (Englewood Cliffs, N. J.: Prentice Hall).

Pyatetskii-Shapiro, I. I. and M. L. Shik. (1964) Spinal regulation of movement. Biophys. 9, 525-530.

Reaves, J. M. (1973) The "coalition": A reaction to the machine metatheory in cognitive psychology. Unpublished manuscript, Center for Research in Human Learning, University of Minnesota.

Requin, J., M. Bonnet, and M. Granjon. (1968) Evolution du niveau d'exitabilité médullaire chez l'Homme au cours de la periode préparatoire au temps de réaction. Journal de Physiologie 1, 293-294.

Requin, J. and J. Paillard. (1971) Depression of spinal monosynaptic reflexes as a specific aspect of preparatory motor set in visual reaction time. In Visual Information Processing and Control of Motor Activity. (Sofia: Bulgarian Academy of Sciences), pp. 391-396.

Roberts, T. D. M. (1967) Neurophysiology of Postural Mechanisms. (London: Butterworths).

Roeder, K. and A. Treat. (1961) The reception of bat cries by the tympanic organ of Noctuid moths. In Sensory Communications, ed. by W. A. Rosenblith. (Cambridge, Mass.: MIT Press).

Rowell, C. H. F. (1964) Central control of an insect segmental reflex. I. Inhibition by different parts of the central nervous system. J. Exp. Biol. 41, 559-572.

Ruch, T. C. (1965a) Transection of the human spinal cord: The nature of higher control. In Physiology and Biophysics, ed. by T. C. Ruch and H. D. Patton. (Philadelphia: W. B. Saunders).

Ruch, T. C. (1965b) Pontobulbar control of posture and orientation in space. In Physiology and Biophysics, ed. by T. C. Ruch and H. D. Patton. (Philadelphia: W. B. Saunders).

Schane, S. A. (1973) Generative Phonology. (Englewood Cliffs, N. J.: Prentice-Hall).

Schneider, G. E. (1969) Two visual systems. Science 163, 895-902.

Shaw, R. E. (1971) Cognition, simulation and the problem of complexity. J. Struct. Learn. 2, 31-44.

Shik, M. L. and G. N. Orlovskii. (1965) Coordination of the limbs during running of the dog. Biophys. 10, 1148-1159.

Shik, M. L., G. N. Orlovskii, and F. V. Severin. (1966) Organization of locomotor synergism. Biophys. 11, 1011-1019.

Shik, M. L., F. V. Severin, and G. N. Orlovskii. (1966) Control of walking and running by means of electrical stimulation of the mid-brain. Biophys. 11, 756-765.

Sperry, R. W. (1952) Neurology and the mid-brain problem. Amer. Sci. 40, 291-312.

Surguladze, T. D. (1972) Functional changes in the segmental apparatus of the spinal cord on execution by man of rhythmic movements. Biophys. 17, 141-145.

Sutherland, N. S. (1973) Intelligent picture processing. Paper presented at Conference on the Evolution of the Nervous System and Behavior, Florida State University, Tallahassee.

Sverdlov, S. M. and Ye. V. Maksimova. (1965) Inhibitory influences of efferent pulses on the motor effect of pyramidal stimulation. Biophys. 10, 177-179.

Taub, E. and A. J. Berman. (1968) Movement and learning in the absence of sensory feedback. In The Neurophysiology of Spatially Oriented Behavior, ed. by S. J. Freedman. (Homewood, Ill.: Dorsey Press).

Tinbergen, N. (1951) The Study of Instinct. (Oxford: Clarendon Press).

Trevarthen, C. B. (1968) Two mechanisms of vision in primates. Psychologische Forschung 31, 299-337.

Turvey, M. T. (1973) On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. Psychol. Rev. 80, 1-52.

Turvey, M. T. (1974) A note on the relation between action and perception. In Psychology of Motor Behavior and Sports, ed. by M. Wade and R. Martens. (Urbana, Ill.: Human Kinetics).

Veber, H. V., I. M. Rodionov, and M. L. Shik. (1965) "Escape" of the spinal cord from supraspinal influences. Biophys. 10, 368-370.

von Foerster, H. (1960) On self-organizing systems and their environments. In Self-organizing Systems, ed. by M. C. Yovits and S. Cameron. (New York: Pergamon Press).

Weimer, W. B. (1973) Psycholinguistics and Plato's paradoxes of the Meno. Amer. Psychol. 28, 15-33.

Weiss-Fogh, T. (1964) Control of basic movements in flying insects. In Homeostasis and Feedback Mechanisms, Symposia of the Society for Experimental Biology, No. 18. (Cambridge: Cambridge University Press).

Wiersma, C. A. (1962) The organization of the arthropod central nervous system. Amer. Zool. 2, 67-68.

Wilson, D. M. (1962) Bifunctional muscles in the thorax of grasshoppers. J. Exp. Biol. 39, 669-677.

Zinkin, N. I. (1968) Mechanisms of Speech. (The Hague: Mouton).

Two Questions in Dichotic Listening*

Michael Studdert-Kennedy[+]
Haskins Laboratories, New Haven, Conn.

The first question concerns the mechanism of perceptual asymmetries. Most investigators have accepted Kimura's (1961a, 1961b) proposal that these asymmetries reflect the asymmetric functions of the cerebral hemispheres. There is, in fact, so much evidence in favor of this hypothesis that it would be difficult to do otherwise. However, not everyone has accepted her structural account of each input's privileged access to its contralateral hemisphere. Kimura (1961a, 1961b, 1967) attributed this privileged access to functional prepotency of contralateral over ipsilateral ear-to-hemisphere connections. Contralateral prepotency rested on the greater number of contralateral than of ipsilateral connections, combined with afferent and perhaps central occlusion of the ipsilateral connections during dichotic competition. Occlusion is evidently not essential, since a sensitive measure of lateralization, such as reaction time, may reveal monaural ear advantages even on quite simple tasks (e.g., Haydon and Spellacy, 1973; Fry, 1974; Morais and Darwin, 1974). However, there is strong evidence from work on split-brain patients that dichotic competition does induce occlusion. Milner, Taylor, and Sperry (1968), Sparks and Geschwind (1968), and, more recently, Zaidel (1973, 1974) have demonstrated that, while these subjects perform equally with left and right ears on monaural identification of digits or nonsense syllables, their dichotic performance reveals a massive, often total, left-ear loss. Moreover, Zaidel (1974) has evidence pinpointing the locus of occlusion as central rather than subcortical. These investigators interpreted their results to make explicit what had been implicit in Kimura's original model, namely, that when normal right-handed subjects attempt to recognize the left-ear input of a dichotically presented pair, they do so from a "degraded" signal that has traversed an indirect path from left ear to right hemisphere, and from right hemisphere to left hemisphere across the corpus callosum.

Central to this account is the assumption of total asymmetry of perceptual function. Variations in ear advantages across phonetic classes (stop consonants, glides, vowels) (Shankweiler and Studdert-Kennedy, 1967; Studdert-Kennedy and Shankweiler, 1970; Cutting, 1974b) would not, according to this model, reflect variations in the degree to which the two hemispheres are engaged in their processing but variations in the degree to which the phonetic classes are liable to transcallosal degradation. Furthermore, as the model would predict, these variations can be eliminated, and the vowels induced to yield a right-ear advantage, if their relative clarity is reduced by presenting them at lower signal-to-noise ratios (Weiss and House, 1973), or as members of an acoustically confusable stimulus set (Godfrey, 1974; cf. Darwin and Baddeley, 1974). Similarly, variations in ear advantages across individuals would not reflect variations in degree of hemispheric asymmetry but variations in degree of contralateral prepotency (Shankweiler and Studdert-Kennedy, 1975). Again, as the model would predict, contralateral prepotency can be eliminated, if the relative clarity of the right-ear input is reduced by presenting it either at a lower signal-to-noise ratio or appropriately filtered (Cullen, Thompson, Hughes, Berlin, and Samson, 1974). Furthermore, individual ear advantages for signals of matched intensity are highly correlated with the amount of which right-ear signal intensity must be reduced in order to eliminate its advantage (Brady-Wood and Shankweiler, 1973). In short, Kimura's model has been widely accepted, and makes sense of a good deal of data.

---

*This paper will appear as an editorial in the second of two special issues of Brain and Language (vol. 2, no. 2, April 1975), devoted to dichotic studies and edited by the author.

[+] Also Queens College and the Graduate Center of the City University of New York.

Nonetheless, two recent studies report results that are incompatible with a simple wiring account of ear advantages in terms of ear-to-hemisphere connections. First, Goldstein and Lackner (1974) have demonstrated that laterality effects may be influenced by subjects' perceived spatial orientations: the normal right-ear advantage for consonant-vowel syllables is reduced if subjects wear prisms that displace their visual environments to the left; it is increased if the prisms displace the environments to the right. Second, Morais and Bertelson (1973) have shown that the strongest perceptual advantage accrues to sounds originating in the median plane, the direction of gaze: although subjects display a significant right-speaker advantage for competing consonant-vowel (CV) syllables presented over left and right loudspeakers, they show a significant front-speaker advantage if the syllables are presented over either front and left or front and right loudspeakers. Both these studies implicate localization mechanisms and suggest that the routing of signals to hemispheres rests, at least in part, on some low-level decision as to the spatial origins of the signals. Evidently, whatever factors determine perceived localization (including, presumably, relative intensity, temporal relations between incoming signals, attention and ear-to-hemisphere connections) will determine the proportion of incoming information that is routed to one or another of the hemispheres. The relative degrees of contralateral/ipsilateral ear-to-hemisphere connections would then have their effect on ear advantages indirectly, as by-products of their roles in auditory localization (cf. Haggard, in press).

These studies are, in fact, more readily compatible with an account of lateral asymmetries in terms of hemispheric specialization and selective attention, or expectancy. Kinsbourne (1970, 1973) first formulated this position, and has elaborated it largely on the basis of visual field studies. He takes as his starting point the fact that each hemisphere serves the contralateral half of space. He proposes that activation of one hemisphere turns attention toward the opposite side, and, at the same time, by the principle of reciprocal innervation, inhibits activation of the other hemisphere. He has demonstrated experimentally, by asking questions that call for either verbal (left-hemisphere) or spatial (right-hemisphere) responses, that subjects orient their gazes away from the midline in a direction contralateral to the putatively activated hemisphere. He has demonstrated, further, that subjects, called upon to retain a list of six words in memory (left-hemisphere activation), while carrying out a tachistoscopic detection or recognition task, display a right-field advantage, where they had previously displayed none, while subjects called upon to rehearse a melody (right-hemisphere activation) display a left-field advantage. Kimura herself (Kimura and Durnford, 1974) has shown that subjects display a right-field advantage for recognition of tachistoscopically presented geometric figures, if they have just performed a similar task for letters, but no advantage, if they do the tasks in reverse order. From here it is a short step for Kinsbourne (1970, 1973) to propose (without necessarily denying central occlusion of the ipsilateral signal) that, given hemispheric specialization as a basis, lateral asymmetries may arise from attentional set induced by the nature of the task rather than from structurally determined contralateral prepotency and transcallosal degradation.

There is, in fact, a lot of evidence that involuntary attention plays a role in determining ear advantages. For example, subjects have difficulty in reversing the "natural" attention of the left hemisphere during a verbal shadowing task: information from the unattended right ear is more likely to intrude than information from the unattended left ear (Treisman and Geffen, 1968). Similar results were reported by Kirstein and Shankweiler (1969) for subjects taking a standard CV syllable test under conditions of directed attention. Furthermore, several studies have shown that dichotically presented vowels, for which a null ear advantage is typical, will yield a right-ear advantage if they are presented in an appropriately biasing experimental context (Spellacy and Blumstein, 1970; Darwin, 1971; Haggard, 1971; Tsunoda, 1975). In short, an attentional model can account for a variety of data that is not readily accommodated by a structural model. But, as Kinsbourne (1973:252) has remarked, what is needed to discriminate between them is an experiment in which materials known to yield a left-ear advantage (e.g., melodies) are mixed with materials known to yield a right-ear advantage (e.g., CV syllables) in the same test. Kimura's model would then predict the usual ear advantages, Kinsbourne's their reduction.

Let us turn now to the second question: the nature and extent of the language hemisphere's peculiar functions. Here, Kinsbourne's model has the advantage that it can accommodate linguistic functions for which asymmetry is partial as readily as those for which asymmetry is total, since the model postulates that the minor hemisphere may be inert owing either to total incapacity or to inhibition by the dominant hemisphere. This is a virtue of the model, since the evidence to date suggests that normal language function entails various processes, some of which are entirely peculiar to the language hemisphere, others of which may, under certain circumstances, be carried out by either hemisphere.

Among the grounds for this statement are the results of work with split-brain patients. The right hemispheres of such patients, although largely mute, have been shown to be capable of considerable verbal comprehension (Gazzaniga and Sperry, 1967; Sperry and Gazzaniga, 1967; Gazzaniga, 1970), including that of complex syntactic and semantic structures (Zaidel, 1973). There are thus important linguistic functions that both hemispheres are equipped to perform. At the same time, as we have seen, the right hemispheres of split-brain patients are almost totally incapable of extracting phonetic information from the left-ear (right-hemisphere) member of dichotically presented digits (Milner, Taylor, and Sperry, 1968; Sparks and Geschwind, 1968) or nonsense syllables (Zaidel, 1974). It was in response to this paradox that Studdert-Kennedy and Shankweiler

(1970:590) proposed that, for the split-brain patient, "...right hemisphere... comprehension rested on auditory analysis which, by repeated association with the outcome of subsequent linguistic processing, had come to control simple discriminative responses."

Essentially the same conclusion has been reached by Zaidel (1973, 1974) on the basis of extensive dichotic studies, and by Levy (1974) on the basis of a series of visual field studies, with split-brain patients. Levy, for example, showed that while the right hemispheres of these patients were able to name pictures of simple, familiar objects (rose, eye, bee), they were unable to recognize that the names of these pictured objects rhymed with "toes," "pie," and "key." In other words, the right hemispheres were able to recognize semantic, but not phonetic, relations. From this and other studies, Levy (1974:161) has concluded that "...there is no evidence whatsoever that the right hemisphere can analyze a spoken input into its phonetic components...." Rather, "...it seems probable that the right hemisphere can decode written or spoken input by having integrated graphologies and phonologies which are tied to their appropriate meanings...and merely utilizes its few whole phonologies to translate input to meaning and meaning to output."

If this is so, then we may further conclude with Studdert-Kennedy and Shankweiler (1970:590) that "to the dominant hemisphere [belongs] that portion of the perceptual process which is truly linguistic: the separation and sorting of a complex of auditory parameters into phonological features." There is, to be sure, scattered evidence that specialization of the language hemisphere may extend as far down into the perceptual process as the detection of characteristic acoustic properties, including temporal order (e.g., Halperin, Nachshon, and Carmon, 1973; Cutting, 1974a, 1974b; Papçun, Krashen, Terbeek, Remington, and Harshman, 1974). However, acoustic analysis does not proceed in isolation. Biological selection of acoustic properties for specialized processing may well have been guided by the function of those properties in determining phonetic structure (cf. Studdert-Kennedy, in press). And, in fact, the mere presence of apt acoustic properties in a speech signal is not sufficient to engage the language hemisphere: for example, recognition of the emotional tone of an utterance, despite its phonetic carrier, engages the right rather than the left hemisphere (Haggard and Parkinson, 1971). Thus, whatever specialized semantic-syntactic processes may subsequently be involved (Zurif, 1974), initial activation of the language hemisphere by speech seems to entail analysis of the signal into its segmental phonetic components. Wood's (1975) elegant work with electroencephalography has lent strong support to this conclusion.

Certainly, phonological analysis may be no more than an instance of a general left-hemisphere cognitive capacity for detailed temporal analysis and abstraction, as compared with that of the right hemisphere for spatial analysis and holistic figure recognition (Bever and Chiarello, 1974; Levy, 1974). Certainly, too, phonological analysis may not be the sole linguistic process to be grounded in such a general capacity: as Zurif (1974) has pointed out, we are sorely in need of well-designed dichotic studies to tease out and identify the semantic-syntactic processes of language perception. Nonetheless, it may be salutary to recall that the single most distinctive property of language as a medium of communication is its construction of meaning from a foundation of meaningless elements (Hockett, 1958; cf. Kimura, in press). Perhaps research will most profitably proceed from the bottom up.

REFERENCES

Bever, T. G. and R. J. Chiarello. (1974) Cerebral dominance in musicians and nonmusicians. Science *185*, 537-539.

Brady-Wood, S. and D. P. Shankweiler. (1973) Effects of amplitude variation on an auditory rivalry task: Implications concerning the mechanism of perceptual asymmetries. Haskins Laboratories Status Report on Speech Research *SR-34*, 119-126.

Cullen, J. K., Jr., C. L. Thompson, L. F. Hughes, C. I. Berlin, and D. S. Samson. (1974) The effects of varied acoustic parameters on performance in dichotic speech perception tasks. Brain Lang. *1*, 307-322.

Cutting, J. E. (1974a) Different speech-processing mechanisms can be reflected in the results of discrimination and dichotic listening tasks. Brain Lang. *1*, 363-374.

Cutting, J. E. (1974b) Two left-hemisphere mechanisms in speech perception. Percept. Psychophys. *16*, 601-612.

Darwin, C. J. (1971) Ear differences in the recall of fricatives and vowels. Quart. J. Exp. Psychol. *23*, 46-62.

Darwin, C. J. and A. D. Baddeley. (1974) Acoustic memory and the perception of speech. Cog. Psychol. *6*, 41-60.

Fry, D. B. (1974) Right ear advantage for speech presented monaurally. Lang. Speech *17*, 142-151.

Gazzaniga, M. S. (1970) *The Bisected Brain*. (New York: Appleton-Century-Crofts).

Gazzaniga, M. S. and R. W. Sperry. (1967) Language after section of the cerebral commissures. Brain *90*, 131-148.

Godfrey, J. J. (1974) Perceptual difficulty and the right ear advantage for vowels. Brain Lang. *1*, 323-336.

Goldstein, L. and J. R. Lackner. (1974) Sideways look at dichotic listening. J. Acoust. Soc. Amer., Suppl. *55*, S10(A).

Haggard, M. P. (1971) Encoding and the REA for speech signals. Quart. J. Exp. Psychol. *23*, 34-45.

Haggard, M. P. (in press) Dichotic listening. In *Handbook of Sensory Physiology*, ed. by H. L. Teuber, R. Held, and H. Leibowitz. (New York: Springer-Verlag), vol. 8.

Haggard, M. P. and A. M. Parkinson. (1971) Stimulus and task factors as determinants of ear advantages. Quart. J. Exp. Psychol. *23*, 168-177.

Halperin, Y., I. Nachshon, and A. Carmon. (1973) Shift of ear superiority in dichotic listening to temporally patterned verbal stimuli. J. Acoust. Soc. Amer. *53*, 46-50.

Haydon, S. P. and F. J. Spellacy. (1973) Monarual reaction time asymmetries for speech and non-speech sounds. Cortex *9*, 288-294.

Hockett, C. F. (1958) *A Course in Modern Linguistics*. (New York: MacMillan).

Kimura, D. (1961a) Some effects of temporal-lobe damage on auditory perception. Canad. J. Psychol. *15*, 156-165.

Kimura, D. (1961b) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. *15*, 166-171.

Kimura, D. (1967) Functional asymmetry of the brain in dichotic listening. Cortex *3*, 163-178.

Kimura, D. (in press) The neural basis of language qua gesture. In *Studies in Neurolinguistics*, ed. by H. Avakian-Whitaker and H. A. Whitaker. (New York: Academic Press).

Kimura, D. and M. Durnford. (1974) Normal studies on the function of the right hemisphere in vision. In *Hemisphere Function in the Human Brain*, ed. by S. J. Dimond and J. G. Beaumont. [London: Paul Elek (Scientific Books)], pp. 25-47.

Kinsbourne, M. (1970) The cerebral basis of lateral asymmetries in attention. In *Attention and Performance*, ed. by A. F. Sanders. (Amsterdam: North Holland) *3*, 193-201.

Kinsbourne, M. (1973) The control of attention by interaction between the cerebral hemispheres. In *Attention and Performance*, ed. by S. Kornblum. (New York: Academic Press), vol. 4.

Kirstein, E. and D. P. Shankweiler. (1969) Selective listening for dichotically presented consonants and vowels. Haskins Laboratories Status Report on Speech Research *SR-17/18*, 133-141.

Levy, J. (1974) Psychobiological implications of bilateral asymmetry. In *Hemisphere Function in the Human Brain*, ed. by S. J. Dimond and J. G. Beaumont. [London: Paul Elek (Scientific Books)], pp. 121-183.

Milner, B., L. Taylor, and R. W. Sperry. (1968) Lateralized suppression of dichotically-presented digits after commissural section in man. Science *161*, 184-185.

Morais, J. and P. Bertelson. (1973) Laterality effects in diotic listening. Perception *2*, 107-111.

Morais, J. and C. J. Darwin. (1974) Ear differences for same-different reaction times to monaurally presented speech. Brain Lang. *1*, 383-390.

Papcun, G., S. Krashen, D. Terbeek, R. Remington, and R. Harshman. (1974) Is the left hemisphere specialized for speech, language and/or something else? J. Acoust. Soc. Amer. *55*, 319-327.

Shankweiler, D. P. and M. Studdert-Kennedy. (1967) Identification of consonants and vowels presented to left and right ears. Quart. J. Exp. Psychol. *19*, 59-63.

Shankweiler, D. P. and M. Studdert-Kennedy. (1975) A continuum of lateralization for speech perception? Brain Lang., *2*, 212-225.

Sparks, R. and N. Geschwind. (1968) Dichotic listening in man after section of neocortical commissures. Cortex *4*, 3-16.

Spellacy, F. and S. Blumstein. (1970) The influence of language set on ear preference in phoneme recognition. Cortex *6*, 430-439.

Sperry, R. W. and M. S. Gazzaniga. (1967) Language following surgical disconnection of the hemispheres. In *Brain Mechanisms Underlying Speech and Language*, ed. by C. H. Millikan and F. L. Darley. (New York: Grune and Stratton), pp. 108-121.

Studdert-Kennedy, M. (in press) Speech perception. In *Contemporary Issues in Experimental Phonetics*, ed. by N. J. Lass. (Springfield, Ill.: C. C Thomas).

Studdert-Kennedy, M. and D. P. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. *48*, 579-594.

Treisman, A. and G. Geffen. (1968) Selective attention and cerebral dominance in perceiving and responding to speech messages. Quart. J. Exp. Psychol. *20*, 139-150.

Tsunoda, T. (1975) Functional differences between right- and left-cerebral hemispheres detected by the key-tapping method. Brain Lang. *2*, 152-170.

Weiss, M. S. and A. S. House. (1973) Perception of dichotically presented vowels. J. Acoust. Soc. Amer. *53*, 51-58.

Wood, C. C. (1975) Auditory and phonetic levels of processing in speech perception: Neurophysiological and information-processing analyses. J. Exp. Psychol.: Human Perception and Performance *104*, 1-33.

Zaidel, E. (1973) Linguistic competence and related functions in the right hemisphere of man following cerebral commissurotomy and hemispherectomy. Unpublished Ph.D. thesis, California Institute of Technology.

Zaidel, E. (1974) Language, dichotic listening, and the disconnected hemispheres. Paper presented at the 15th annual meeting of the Psychonomic Society, November 23-26, Boston, Mass.

Zurif, E. B. (1974) Auditory lateralization: Prosodic and syntactic factors. Brain Lang. 1, 391-404.

On the Relationship of Speech to Language*

James E. Cutting[+] and James F. Kavanagh[++]

At first glance the phrase speech and language may appear redundant. Just as with the phrase null and void, laypersons and scientists alike often view the terms as duplications of one another. At second glance one realizes that this is not true: speech could be considered as the spoken vehicle of language. This view would seem to place speech inside language, giving it the same relationship as the part to the whole.

Only recently have speech scientists, psychologists, linguists, anthropologists, and philosophers, among others, begun to look in earnest beyond these first and second glances; only recently have they begun to treat speech and language as separate entities in a symbiotic partnership. This third view, just as the previous ones, may not be entirely correct, but it has considerable intuitive and empirical support. Moreover, it provokes some interesting questions. For example, if language and speech are independent, it must be possible to have language without speech and speech without language.

LANGUAGE WITHOUT SPEECH

There are a number of contenders for the label "language without speech." Many are controversial. Consider first the sign languages of the deaf, particularly American Sign Language (ASL). This mode of communication uses hand gestures in relationship to the head and torso, along with large doses of eye contact, to convey meaning from signer to sign-receiver. Clearly, there is no speech in ASL, no tongue movements to shape sound. This, among other features of sign languages, has led some researchers to question whether ASL is, indeed, a language at all. The title of Hans Furth's book, Thinking Without Language, bespeaks this position; Bellugi and Klima's forthcoming book, The Signs of Language, on the other hand, will have a different view. Rather than enter into this debate, which may be more acrimonious than fruitful, some have chosen to observe how sign languages differ from spoken languages. We shall return to these observations in some detail.

---

*To be published in The Journal of the American Speech and Hearing Association.

[+] Wesleyan University, Middletown, Conn., and Haskins Laboratories, New Haven, Conn.

[++] National Institute of Child Health and Human Development, Washington, D.C.

[HASKINS LABORATORIES: Status Report on Speech Research SR-41 (1975)]

Another illustration of language without speech is seen in certain cases of congenital anarthria, where the patient never acquires the ability to speak but can understand language easily. Christy Brown, for example, grew up with little speech, but had language abilities refined enough to write the best seller _All the Down Days_. In an even more extreme example, Lenneberg (1962) reports the case of a child who had _no_ speech, but could understand language nearly as well as his unafflicted agemates.

A third possibility of language without speech is the most controversial, and concerns the considerable efforts undertaken to teach language to chimpanzees. It is clear that chimps cannot learn to talk even given the most extensive training: their vocal tract simply appears to be inadequate (Lieberman, Crelin, and Klatt, 1972). They can, however, become remarkably adept at using the sign gestures of ASL (Gardiner and Gardiner, 1969; Fouts, 1973), at manipulating plastic symbols on a magnetized board to convey meaning (Premack, 1971), or at "reading and sentence completion" of computer-displayed geometric symbols (Rumbaugh, Gill, and von Glaserfeld, 1973). Are chimps capable of language behavior, or merely languagelike behavior? Fodor, Bever, and Garrett (1974) remain unconvinced that these demonstrations are even relevant to language; Lieberman (1973), on the other hand, finds them compelling. This is another controversy that we choose to avoid. Regardless of whether chimps do or do not have language, we think it useful to observe what chimpanzees can and cannot do for the purpose of investigating the scope of language without speech.

SPEECH WITHOUT LANGUAGE

There are also several contenders for the label "speech without language." Again, some are controversial. The early babbling of the infant is often thought to be nonlinguistic (Jakobson, 1968; Kewley-Port and Preston, 1974); brain-damaged patients with extreme forms of expressive aphasia often speak with good rhythm and intonation patterns, but with no apparent words or meaning (Green, 1973); and the "speaking in tongues," or glossolalia, often associated with Pentecostal churches, has been found to lack underlying structures necessary in more worldly languages (Samarin, 1972). Some consider all three of these examples more akin to song than to language, and, indeed, the derivation of the word _glossolalia_ (from the Greek words _glosso_ meaning tongue, and _lalia_, meaning lullaby) seems to support this view. One can avoid any controversy, however, by looking to song lyrics themselves for examples of speech without language. Surely, all critics agree that the "fa-la-la-la-la" of certain Christmas carols and the "sha-boom sha-boom" of certain popular songs of the 1950s and 1960s lack linguistic content. These are speech sounds for sound's sake. They have no duality of patterning so familiar to spoken languages (Hockett and Altmann, 1958); that is, they are sound without meaning.

A FRAMEWORK FOR THE STUDY OF SPEECH AND LANGUAGE

If speech and language are as isolable from one another as they appear to be in the above examples, a number of interesting questions arise. How do speech and language function in concert, and, more particularly, what are the effects of one on the other? In October 1973, a group of researchers, many of whom are directly involved in the controversies mentioned earlier, met under sponsorship of the National Institute of Child Health and Human Development at Columbia, Maryland, for three days of presentations and discussions. Their topic was the role of speech _in_ language.[1] Alvin Liberman, who introduced the conference, noted that the underlying question that motivated the meeting was not an established one: Can we increase our understanding of language when we take into account that it is spoken? In other words, in this allegedly symbiotic partnership, what are the effects of speech on language? Most of the participants had not previously addressed themselves to this query, but rather to research questions related to it in areas such as speech production, oral biology, speech perception, phonology, syntax, animal communication, sign languages of the deaf, language evolution, and symbolic processes.

A framework helpful in assessing the role of speech in language is to consider the output "terminals" of the communication chain in man: intellect and vocal tract, or, more simply, mind and mouth. In this communication chain, imagine the intellect as the initiating terminal and ultimately as the receiving terminal in the communication process; the vocal tract and the ear are the proximal output and input terminals.[2] Keeping this framework in mind, one can think of the rules of language as the interface mechanism (or "grammar" as linguists would call it) between intellect and the lower way stations in the chain. Likewise, one can view the rules of speech as the grammar between the vocal tract and the higher mechanisms of the chain. In this manner, speech and language are seen as different rule systems working at different levels. More specifically, there are the _phonological_ rules of speech, and the _semantactic_ rules of language. This latter term is a combination of the more familiar terms _semantic_ and _syntactic_ as used by Ross at the conference.

Given the framework outlined thus far, there may appear to be a gap in the system. What, for instance, is the interface between the grammars of speech and language? The answer appears to be that there is none: they interact directly with one another. Interaction implies mutual adjustments and mutual change. Thus, a logical extension of this model is that speech works upward in the communication chain to constrain and alter language, and perhaps even intellect; language, working in the reverse direction, exerts downward constraints to alter speech, the vocal tract, and perhaps the ear as well. Evidence for evolutionary change in the shape of the mind is difficult to come by. Evidence for evolutionary change in the shape of the vocal tract, however, can be seen by comparing fossil skulls of certain homonids with those of modern man. Philip Lieberman, at the conference and in previous publications, suggested that the human vocal tract assumed its present configuration specifically to make speech possible. This view is contrary to the more venerated notion that speech is merely a faculty overlaid on eating and respiratory functions. Evidence that the newer, evolutionary view is correct stems partly from the fact that man, in addition to being the only creature to speak, may be the only creature to choke easily on his food. While these downward constraints on the vocal tract are important, it is the upward constraints, those that shape language and the mind, that are perhaps the more interesting changes in evolution, and it is those that are more directly relevant to the role of speech in language.

---

[1] The conference was entitled "Communication by Language: The Role of Speech in Language." Those who attended or contributed to the conference included, in addition to the present authors, Ursula Bellugi, James F. Bosma, Peter D. Eimas, Jerry A. Fodor, Gordon W. Hewes, Ira J. Hirsh, Janellen Huttenlocher, James J. Jenkins, R. Paul Kiparsky, Edward S. Klima, Alvin M. Liberman (co-chairman with Kavanagh), Philip Lieberman, Peter Marler, Ignatius G. Mattingly, David S. Palermo, David Premack, Peter C. Reynolds, John Robert Ross, Robert E. Shaw, William C. Stokoe, Jr., and Michael Studdert-Kennedy. The conference proceedings are published as _The Role of Speech in Language_ (Kavanagh and Cutting, 1975).

[2] We have purposefully borrowed from Denes and Pinson (1973) the notion of a _speech chain_--which includes the vocal tract, air vibrations, and the ear--and extended it to include intellect at both ends. The result could still be called the speech chain, but we propose to substitute the hands and eye for the vocal tract and ear, respectively, when dealing with sign language, and to substitute for the human intellect that of chimpanzees and even birds when dealing with animal communication. The end result can only be considered the _communication chain_.

Three approaches seem relevant to our goal of understanding the relationship of speech to language. First, one can focus on speech itself, or more specifically on phonology, to obtain insights about the workings of language and of the mind. Second, one can trace inasmuch as possible the development of speech in man and child, making inferences about language and intellect behind the expansion of ability in vocal communication. Third, one can look at the linguistic structures of sign language, the most important form of language without speech, with an eye toward differences between sign and speech and how they affect the more abstract levels of the communication chain.

PHONOLOGY AND THE LANGUAGE OF THE MIND

Speech scientists and linguists have always treated speech and language as separate entities. Their problem, as Paul Kiparsky and John Robert Ross told the conference, is a failure to map out, in a nontrivial manner, the functional and structural relationship between them. One way to accomplish this appears to be to observe interactions of phonology and semantax. For example, *John's in Boston* is a perfectly good sentence. *Bill's happier in Portland than John's in Boston*, however, is not. In this example by Kiparsky, the phonology of the phrase *John is in Boston* is dictated by higher-level rules--mind shapes mouth. Are there examples of mouth shaping mind, where phonological rules dictate semantactic structure? Perhaps, but they appear much more difficult to find at present.

A second way to accomplish our goal, then, is to draw parallels between phonological and semantactic grammars. Ross outlined several, one of which might be termed a simplification process at both levels. At the semantactic level speakers tend to reduce complex sentences to simple ones. Rather than saying *I know someone who is tall*, for example, one is more likely to say a shorter and simpler sentence *I know someone tall*. At the phonological level speakers tend to reduce multisyllable utterances into one or two syllable utterances, especially when among friends. Thus, *did you eat yet?* is easily shortened to *did y'eat yet?* and finally *j'eat yet?* There are, however, problems with such parallels. Just as correlation does not imply causation in statistical analysis, parallels between phonology and semantax do not necessarily imply upward or downward constraints in the communication chain. Nevertheless, such groundwork is vital to the field if it is to become ripe for new discoveries.

DEVELOPMENT OF SPEECH IN MAN AND CHILD

We can only sketch some of the more important and interesting issues in this awesomely broad, second approach. One issue, for example, is why speech developed so late in man--perhaps only 50,000 years ago--and develops so late in the child--between one and two years. One reason for this "lateness" is directly related to functional anatomy, as suggested earlier. Lieberman reconstructed from fossil remains the vocal tracts of premodern man and compared them to those of modern adults and neonates. Of the three, the vocal tracts of premodern man and the modern neonate were most similar and lacked the particular shape requisite for full-range speech sounds of the modern adult. Thus, ontogeny recapitulates phylogeny, and one reason for the "late" development of speech both in man and in the individual child appears to be physiological inadequacy. Physiology, however, cannot be the entire answer. The child's vocal tract becomes adequate many months before speech is produced in a regular fashion. By inference, this may have been true for premodern man as well. Therefore, other factors such as cognitive ability must be considered: men and children need something to say as well as the apparatus to say it with.

The tardiness yet pervasiveness of speech seems paradoxical. Whereas language without speech is thought by some to be impoverished, language abilities may develop before speech abilities. Gordon Hewes (1973), for example, has suggested that language first developed in prehistory through the use of gestures perhaps similar to those of modern sign languages; and William Stokoe, at the conference, claimed that sign language develops in the deaf child before speech develops in the normal child. These notions, if true, would seem to indicate that sign is more "natural" to language than is speech--an irony indeed. The resolution of this apparent paradox may be to assume that speech and language evolved separately, perhaps at separate times, and only later coevolved into a more or less unified and symbiotic system. The independent evolution of speech is supported by Mattingly (1972). He noted structural parallels between speech, certainly the most complex signaling system in nature, and various rudimentary animal communication systems, which could hardly be called language or even languagelike.

If language-by-sign developed earlier than speech, or at least independent of it, why did speech supplant sign as the major vehicle of language? Surely the answer must be more complex than to free the hands for manual skills such as hunting, gathering, tool-making, and cooking. One reason, we can safely assume, concerns speed of communication. At the conference, Ursula Bellugi noted that modern sign languages are not as rapid as speech (see also Bellugi and Fischer, 1972). Protosign was surely no faster and could not compete with the more rapid, newly evolved vocal form of communication. This view seems reasonable. Even speech is woefully slow at times. Slips of the tongue often reveal telescopic jumps where speech skips ahead many syllables as if to catch up with the more nimble leaps of the mind. There may be evolutionary and everpresent pressures to speed up communication. Perhaps sign lost out to speech because of them.

Another reason for the change from sign to speech may be related to modality. Put in its simplest form, almost all objects in nature are opaque to the eye, but few are "opaque" to the ear; that is, one cannot see through foliage and rocks, but he or she can hear "through" or at least around them. This feature becomes vitally important when one walks or runs through dense jungles and high grasses--as did man's forebears--where vision is often very restricted. In this light, it is necessary to consider the role of vocalizations in animal communication, comparing them to the role of speech in language. Two types of creatures are of particular interest: primates, because of their evolutionary relationship to man, and songbirds, because of impressive analogs between the acquisition of birdsong and of speech.

Peter Marler told the conference about comparative ethological trends in Asian and African primates that are relevant to development of speech and language in man. As primates develop a more complex vocal repertory, they also tend to become more terrestrial (living on the ground rather than in trees), less territorial, and more inclined to live in large troops. All of these are trends toward the social state of man. More importantly, a major change of emphasis in communication appears to be correlated with this trend. With these other developments, the largest portion of signaling repertories shifts from between-troop warning calls and vocal displays to within-troop social calls. Parallel to this change in type of communication is a change in "vocabulary," from a discrete and limited set of calls to a graded and less bounded call system. This trend allows for a larger and more subtle repertory of vocal sounds. Marler interprets this move toward graded systems as approximations of speechlike behavior in man.

From a view external to that of the speech perceiver, Marler is correct: human speech is extremely graded. For example, if many samples of human speech were displayed on sound spectrograms and compared to each other, one would see an impressive dearth of discrete differences among the speech sounds. They would look, as Hockett (1955) has suggested, like so many smashed Easter eggs. To be sure, humans do not perceive speech in a graded or continuous manner; it seems to segment itself into syllables and phonemes almost automatically. How we accomplish the feat of reassembling the smashed eggs, the units of speech, remains largely a mystery, as those involved in the problem of machine recognition of speech can attest. Viewed from the "outside," then, as any computer or intelligent nonhuman must view speech, it is strikingly graded and continuous. This raises an interesting issue. Just as computers have difficulty segmenting human speech, humans have difficulty segmenting the graded calls of chimpanzees, which are necessarily viewed from the "outside." Do chimps and other primates segment their graded vocalizations? This is an important question. Whether they do or do not, however, the emphasis on the evolutionary role of speech in language might well be placed on perception rather than on production.

The prominence of perception over production receives support from birdsong, as well as from speech itself. Consider first the songs of passerine birds. The white-crowned sparrow, for example, must hear versions of his species-specific song if he is to produce it, and he must hear it during his first year, well before he begins to sing it. Furthermore, he must continue to hear himself and fellow white-crowns as he produces approximations to full song during the following year. Surgical deafening at any time before the advent of full song inhibits the production process and full song will not develop. In an analogous fashion, humans may need to perceive speech before they can start to produce it, and later they may need to compare their productions with those of adults before speech becomes regularized. Critical periods for humans are probably much less inflexible than for songbirds, but a parallel is unmistakable. Evidence suggests that infants can perceive speech-relevant sounds well before they can produce them. Peter Eimas, at the conference and in previous work (Eimas, Siqueland, Jusczyk, and Vigorito, 1971), presented data that one-month-old infants are able to discriminate phonetically relevant features in computer-generated tokens of speech much better than similar but phonetically irrelevant features. These discriminations, which are requisites for speech segmentation, occur at least a year before the same phonetic distinctions will be accurately produced (Kewley-Port and Preston, 1974).

If one considers speech as a "species-specific song" in a broad sense, infants must be exposed to elements of the "full song" long before they can produce it. Infants deaf from birth have extreme difficulty in acquiring speech, but children who become deaf later, at age five or ten, for example, may continue to have remarkably normal speech for the rest of their lives, just as the white-crowned sparrow deafened after the development of song in his second year will continue to sing in a normal manner.

In addition, like humans, white-crowned sparrows have dialects according to geographical region. These aspects of full song appear to be first learned through exposure long before the young bird ever sings. Recent research with humans has shown that young infants begin to learn by the age of two months the more exotic, "dialectic" aspects of their to-be-native language, which two monthers in other lands will not have learned (Streeter, 1974). Again, this is long before the sounds will be produced and used to convey meaning in spoken language.

Ontogenetic and phylogenetic observations about the acquisition of speech have gone well beyond our first approach to the role of speech in language, that of observing phonology itself. Yet, like that approach, this second one is still very new and has only recently begun to bear fruit. Evidence from the calling systems of primates and of songbirds, as well as that presented by Mattingly (1972), supports the view that speech has strong evolutionary ties independent of language. Thus far, however, we have presented little information about how speech as a signaling system was applied to language and what effects that application had. This is crucial to our goal of discerning the role of speech in language. Our third approach is addressed to this question, but necessarily in an indirect fashion.

COMPARISONS OF SIGN LANGUAGE AND SPEECH

If perception is a requisite for production of speech, as we have suggested earlier, what is the effect on language and intellect when that channel of perception is totally blocked? Robbed of audition from birth, the deaf human may have no opportunity to develop speech and may have to use the slower sign-gestures to communicate. Some have suggested that the choice of sign over speech may have intellectual costs. In some cases, however, it is clear that there are no such costs to deaf signers even when that individual is compared to normal speakers. But the question about the size, or intellectual capacity of the mind should be separated from the question about the shape of the mind. The shape of a soundless language and the intellect behind it is the issue addressed by Bellugi, Klima, and Stokoe at the conference.

Aside from the sheer scope of trying to compare all of sign to all of speech, there are several other problems. One is data base. Only one person in a thousand is deaf, and only one deaf person in ten is the child of deaf parents. Thus, it is only one child in ten thousand who learns sign as a native language. The other nine in ten thousand will probably learn sign, but in conjunction with speech, which might "contaminate" the study of pure sign. Second, there is the problem of the pervasive influences of the spoken culture around enclaves of native signers. In America, among signers of ASL, there are at least three forms of signs: (1) finger-spelled words of English, which may not have a direct analog in sign, (2) signed English, which is an approximation of English morphology and syntax, and (3) natural sign. Native signers typically use all three, but it is only the latter that is of primary interest here. Third, there are differences between sign and pantomime, which must be closely observed. Sign is only partially iconic, whereas pantomime is almost exclusively so. The icon, or visual image, is often drawn or shown with the fingers and hands in front of the signer/pantomimist and referred to later in the sign/pantomime discourse. With all these complexities, it becomes evident that any effort to study sign language by the nonsigning researcher is difficult without the aid of native-signing collaborators. Stokoe, at Gallaudet in Washington, D.C., and Bellugi and Klima, at the Salk Institute in California, rely heavily on their deaf colleagues.

Comparing sign to speech, one first finds that sign has no sounds, no phones, and no "phonology" in the normal sense. Phones, or phonemes, are the meaningless units that make up spoken words and sentences: they are the /b/, the /o/, and the /t/ that make up the word boat. Are there such meaningless units in sign? Yes, but they do not correspond exactly to the phoneme or even to the syllable. The three important features of a sign, in a psychological sense, appear to be the hand configuration, the place of articulation of the designating hand with respect to the head, torso, or other hand, and the movement of the hand once it is there. Each configuration, place, and movement is meaningless in itself, just as phones are meaningless. It may seem ironic that meaninglessness is important to communication; one could easily have predicted the opposite. Nevertheless, it is the combination of such units that makes meaningful words and signs possible. Some combinations are easier than others to produce, and some, while easy to articulate, simply seem wrong: whereas bnick (to use an example from Klima) is easy to pronounce, it does not conform to English phonology. Thus, phonological rules constrain the possible combinations of phones. There are sign analogs to bnick. Certain hand configurations seem wrong to native signers when accompanied by certain movements or coupled with a certain place of articulation. In the broadest sense, then, sign has a "phonology" analogous to that of any spoken language. When comparing the influences of sign and of speech on language and intellect, one must remember that he is not comparing systems in the presence or absence of phonology, but rather systems with different phonologies. This makes our task all the more difficult, but all the more intriguing as well.

From this brief look, it may appear that the phonology of sign consists merely of articulatory do's and don'ts. This is incorrect. The phonology of sign, if we may use the phrase, is broader than that. Perhaps more important than articulation rules are the temporal constraints alluded to earlier. Since speech is faster than sign, sign must somehow try to catch up. Bellugi and Fischer (1972) asked the question: How does sign save time and still communicate unambiguously? The answers fall into at least three categories: doing without, incorporation, and bodily or facial shifts. Doing without often means simply omitting the redundancy of spoken and written language. Bellugi and Fischer note that the signed version of the complex sentence John likes Mary, so he goes and *visits her a lot, and he often takes her out to dinner, though sometimes he cooks for her* would scan (when translated back into English) something like: JOHN LIKE MARY, WELL, GO VISIT MUCH, OFTEN TAKE OUT EAT, BUT SOMETIMES COOK FOR. Clearly, much has been dropped in the signed version, but the message is essentially identical. Incorporation, the second way to shortcut in sign, takes many forms. Often sign incorporates iconic spatial referents. A simple example would be to compare the two signed sentences corresponding to *She is bigger than me* and *She is much bigger than me*. Both signed sentences would take the same amount of time to "pronounce" but in the second form the sign for *large* (bigger than) would be exaggerated. Bodily and facial shifts, the third major class of sign accelerators, deliver information in parallel with the sign discourse. For example, the hand gestures corresponding to the sentences *I know that* and *I don't know that* are identical. The signed version of the second sentence is accompanied by a headshake, or a small frown, indicating negation, thus saving time. Bellugi and Fischer do not claim that this small list includes all time-saving devices in sign, but it is interesting that these three--doing without, incorporation, and bodily shifts--are exactly those that make face-to-face *verbal* communication so much easier and faster than communication by telephone. Furthermore, they are exactly the reasons that conferences and meetings, where people are often drawn together from great distance and at great expense, are more prevalent and more rewarding than conference telephone calls, even though the latter may be cheaper.

Systematic comparisons of sign and speech have only just begun. Much of the present research may look like so much dabbling, but underlying it is the need for asking the right questions, which cannot be accomplished until we have dabbled. Promising approaches have been taken by Bellugi, Klima, and others, and a few deserve mention here. First, just as there are slips of the tongue in speech, there are "slips of the hand" in sign. Fromkin (1973) has analyzed these faux pas in speech and found richly rewarding insights into the serial organization of speech. Studies of slips of the hand will be equally rewarding in unraveling the structures of sign. Second, just as there are infantile or "baby talk" forms of speech, there are infantile forms of sign. In some ways these are similar to speech, in others they are different. The acquisition of signs by children is certainly worthy of study to the extent that, for instance, Brown (1973) has studied the first spoken sentences of normal children. Third, psychologists have been interested in the different types of forgetting that occur for information presented by eye and information by ear. Typically, these memory errors are different, particularly with regard to most recently occurring items in a list. Bellugi presented to the conference evidence that signers forget lists of words in a manner nearly identical to the way normal listeners forget lists of words that are spoken, but not in the manner normals forget those words when written. By extension, perceiving sign may be more similar to listening to speech than to reading, even though both sign-receiving and reading are visual skills.

HOLISM OF SPEECH AND LANGUAGE

A word of caution must be inserted at this point. While it is clear that speech and language can be logically separated, whether by comparing phonology and semantax, by postulating their separate genetic developments, or by comparing language with and without speech, they remain part of one system. James Jenkins and Robert Shaw, playing devil's advocates at the conference, saw a danger in the fractionation of speech and language and subsequent overanalyses that may follow. As a historical case in point, they noted how the field of aphasia research has suffered from this very division. After reviewing 50 years of empirical research on large samples of brain-damaged patients, they found few, if any, examples of *pure* productive aphasia (language without speech) or *pure* receptive aphasia (speech without language).

In summary, then, perhaps the third view of the relationship between speech and language, that they are separate entities in a symbiotic partnership, should be tempered. Separateness may imply an independence that surely does not exist in the normal speech-language-communication system in man. Accepting this cautionary note, exploration into the relationship of speech to language has only just begun and should prove a fascinating and fruitful line of research for those in a number of scientific disciplines that converge on communication in man.

REFERENCES

Bellugi, U. and S. Fischer. (1972) A comparison of sign language and spoken language. Cognition 1, 173-200.

Brown, R. (1973) *A First Language: The Early Stages*. (Cambridge, Mass.: Harvard University Press).

Denes, P. B. and E. N. Pinson. (1973) *The Speech Chain*. (Garden City, N. Y.: Anchor Press).

Eimas, P. D., E. R. Siqueland, P. Jusczyk, and J. M. Vigorito. (1971) Speech perception in infants. Science 171, 303-306.

Fodor, J. A., T. G. Bever, and M. F. Garrett. (1974) *The Psychology of Language*. (New York: McGraw-Hill).

Fouts, R. S. (1973) Acquisition and testing of gestural signs in four young chimpanzees. Science 180, 978-980.

Fromkin, V. (1973) Slips of the tongue. Sci. Amer. 229, 110-117.

Gardiner, R. A. and B. T. Gardiner. (1969) Teaching sign language to a chimpanzee. Science 165, 664-672.

Green, E. (1973) Phonological and grammatical aspects of jargon in an aphasic patient: A case study. In *Psycholinguistics and Aphasia*, ed. by H. Goodglass and S. Blumstein. (Baltimore: Johns Hopkins University Press).

Hewes, G. (1973) Primate communication and the gestural origin of language. Curr. Anthropol. 14, 5-24.

Hockett, C. F. (1955) A manual of phonology. Intl. J. Ling., Memoir 11.

Hockett, C. F. and S. A. Altmann. (1958) A note on design features. In *Animal Communication*, ed. by T. A. Sebeok. (Bloomington, Ind.: Indiana University Press).

Jakobson, R. (1968) *Child Language, Aphasia, and Phonological Universals*. (The Hague: Mouton).

Kavanagh, J. F. and J. E. Cutting, eds. (1975) *The Role of Speech in Language*. (Cambridge, Mass.: MIT Press).

Kewley-Port, D. and M. S. Preston. (1974) Early apical stop production: A voice onset time analysis. J. Phonetics 2, 195-210.

Lenneberg, E. (1962) Understanding language without ability to speak: A case report. J. Abnormal Social Psychol. 65, 419-425.

Lieberman, P. (1973) On the evolution of language: A unified view. Cognition 2, 59-95.

Lieberman, P., E. S. Crelin, and D. H. Klatt. (1972) Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. Amer. Anthropol. 74, 287-307.

Mattingly, I. G. (1972) Speech cues and sign stimuli. Amer. Scient. 60, 327-337.

Premack, D. (1971) Language in chimpanzee? Science 172, 808-822.

Rumbaugh, D. M., T. V. Gill, and E. C. von Glasserfeld. (1973) Reading and sentence completion by a chimpanzee (Pan). Science 182, 731-733.

Samarin, W. J. (1972) Glossolalia. Psychol. Today 6, 48-50, 78-79.

Streeter, L. (1974) The effects of linguistic experience on phonetic perception. Unpublished Ph.D. dissertation, Columbia University (Arts and Sciences).

Rise Time in Nonlinguistic Sounds and Models of Speech Perception

James E. Cutting,* Burton S. Rosner,[+] and Christopher F. Foard[++]

Sawtooth wave stimuli differing in rise time yield perceptual effects previously thought unique to stop consonants. The stimuli are identifiable as either plucked or bowed, as if coming from a stringed instrument. After selective adaptation, they demonstrate boundary shifts similar to those found for stop consonants. Moreover, like stops, they are perceived categorically according to the strictest criteria. Unlike many speech sounds, however, they do not yield a right-ear advantage in dichotic listening. These and other results suggest that speech perception may not use newly evolved, unique mechanisms. Instead, the extensive engagement of the patterned engagements of certain older mechanisms during speech perception may be unparalleled.

Psychologists and laymen alike commonly believe that language distinguishes man from other animals. Recently, however, this view has been challenged. Chimpanzees, for example, are remarkably adept at manipulating sign gestures (Gardiner and Gardiner, 1969; Fouts, 1973) or plastic symbols (Premack, 1971) to produce languagelike behavior. A new and more sophisticated position, therefore, suggests that speech, but not language, is unique to man (Lieberman, 1973): man is the only creature with a supralaryngeal vocal tract equipped to make the complex articulatory gestures necessary in speech. If we grant that the evolution of the vocal tract has made flexible speech productions possible (Lieberman, Crelin, and Klatt, 1972; Lieberman, 1973), a new question arises: Have other mechanisms coevolved to make human speech perceptions equally facile?

There are two general approaches for determining any unique properties of human perception of speech. The first is to test nonhuman primates, observing their perceptions of speech stimuli in particular paradigms and comparing the

---

*Wesleyan University, Middletown, Conn., and Haskins Laboratories, New Haven, Conn.

[+] University of Pennsylvania, Philadelphia.

[++] University of Pennsylvania and Haskins Laboratories.

Acknowledgment: We thank David B. Pisoni for his comments, his encouragement, and for the use of his stimuli; Michael Studdert-Kennedy for careful reading of the manuscript; and Ruth S. Day for the use of her equipment.

[HASKINS LABORATORIES: Status Report on Speech Research SR-41 (1975)]

results to those from humans in similar paradigms. The second is to test humans, observing their perceptions of nonlinguistic stimuli as compared with their perceptions of speech. We have followed this second strategy. This paper shows whether several characteristic findings in speech perception--boundary shifts in selective adaptation, categorical perception, and right-ear advantages in dichotic listening--occur for certain nonspeech sounds as well.

Cutting and Rosner (1974) found that sawtooth and sine wave stimuli differing in rise time are categorically perceived according to criteria suggested by Studdert-Kennedy, Liberman, Harris, and Cooper (1970). These stimuli yield relatively quantal identification functions and produce discrimination functions displaying sharp differentiation across the category boundary and near-chance performance within each category. The boundary cannot be explained in terms of presence or absence of clicks in the signals, and the categories are not determined by the learning of labels (Cutting and Rosner, 1974). The present experiments probed the perception of these stimuli in more depth, testing for selective adaptation through the technique of Eimas and Corbit (1973). We also tested for categorical perception using the delayed discrimination procedure of Pisoni (1973), as well as the original paradigm of Liberman, Harris, Hoffman, and Griffith (1957), and compared these results with those obtained for consonants and vowels. Finally, we sought possible ear advantages in dichotic listening, using a method similar to that of Studdert-Kennedy and Shankweiler (1970; see Cutting, 1974a, 1974b). We shall consider the adaptation results first and then the categorical perception and dichotic listening data.

EXPERIMENT I: SELECTIVE ADAPTATION

Because so much recent experimentation has focused on selective adaptation in speech stimuli (see Eimas, Cooper, and Corbit, 1973; Eimas and Corbit, 1973; Ades, 1974a, 1974b; Cooper, 1974, in press), we sought such effects in our nonspeech stimuli.

Method

The musiclike stimuli of Cutting and Rosner (1974) were used. They were generated on the Moog synthesizer at the Presser Electronic Music Studio, University of Pennsylvania, and recorded on audio tape. They were then digitized and stored on a computer disc file using the pulse code modulation system at the Haskins Laboratories (Cooper and Mattingly, 1969). The stimuli consisted of four nine-item arrays: sawtooth and sine wave stimuli each at 294 and 440 Hz, the items within an array differing in rise time by 10-msec steps. Items varied between 1020 and 1100 msec in duration according to rise time (see Cutting and Rosner, 1974, for fuller details). Stimuli with rapid rise times resemble the plucking of a stringed instrument (like a guitar), whereas more slowly attacked items sound like the bowing of a violin. The particular array of items selected to be identified in pre- and postadaptation tests was the 440-Hz sawtooth continuum. Adapting stimuli were the 0- and 80-msec items from each of the four arrays. (Stimuli with 0-msec rise time actually reached peak amplitude in a quarter-cycle.) Thus, there were eight adaptation situations: adaptation within the same continuum, using 0- and 80-msec 440-Hz sawtooth items; adaptation across frequency, using 0- and 80-msec 294-Hz sawtooth items; adaptation across waveform, using 0- and 80-msec 440-Hz sine wave items; and adaptation across both frequency and waveform, using 0- and 80-msec 294-Hz sine wave items.

Eight adaptation tapes were recorded at the Haskins Laboratories using the pulse code modulation system. All eight followed the same pattern. There were 600 msec between repetitions of the adapting stimulus, 2 sec between successive postadaptation identification items, and 5 sec between the end of the identifications and the beginning of the next adaptation sequence. The first adaptation sequence of each tape consisted of 100 repetitions of the adapting stimulus and then seven items to be identified from the nine-item 440-Hz sawtooth continuum. Five subsequent sequences consisted of 50 repetitions of the same adapting stimulus and seven items to be identified. Thus, one run through any tape yielded 42 postadaptation identifications. The number of observations per stimulus item, however, was distributed such that midrange items with rise times 20 through 60 msec were represented twice as often as extreme items with 0-, 10-, 70-, and 80-msec rise time. Since each subject heard each tape twice in succession, the total number of observations per subject for midrange items was 12 each, and for other items 6 each. A preadaptation identification tape of 90 items in random sequence was also recorded: (9 items in the 440-Hz sawtooth array) by (10 observations per item).

Eight University of Pennsylvania undergraduates and graduate students participated in eight adaptation situations, one situation per experimental session. Subjects were all right-handed, native American English speakers with no history of hearing difficulty. They were tested individually listening through matched Telephonics earphones (Model TDH-39) to diotically presented stimuli played on a Revox tape, recorded at 80 db re 20 $\mu N/m^2$. Each session consisted of a preadaptation identification test followed by two passes through an adaptation tape. Sessions were, on the average, 24 hours apart. The order in which subjects participated in the eight situations followed a balanced design. They wrote down P for pluck or B for bow for each identification.

Results

The results of the adaptation studies appear in the four panels of Figure 1. All identification functions are quite quantal, repeating the findings of Cutting and Rosner (1974). Each panel contains two postadaptation functions, one for identifications following adaptation with a pluck stimulus (0-msec rise time) and the other for those following adaptation with a bow stimulus (80-msec rise time). These lie astride the mean preadaptation identification functions for the two experimental situations. The preadaptation functions were combined since there was little difference between them. Nevertheless, small differences obscured by such averaging could affect assessment of the extent of postadaptation boundary shifts. Thus, shift magnitudes for each of the eight adapting conditions are shown in Table 1, along with corresponding Wilcoxon matched-pairs signed-ranks tests. Each shift was measured for each subject by summing the probabilities of pluck responses across stimuli in the postadaptation situation and subtracting from this total the summed probabilities of such responses in the preadaptation situation. Since items within the test continuum differ in rise time by 10-msec steps, the magnitude of the boundary shift was derived by multiplying the mean difference in probabilities for the 8 subjects by 10.

Six of the eight adaptation situations yielded significant boundary shifts, and seven were in the predicted direction toward the rise time value of the adapting stimulus. Although pre- and postadaptation results were not always significantly different, the boundary shifts shown in Table 1 were great enough

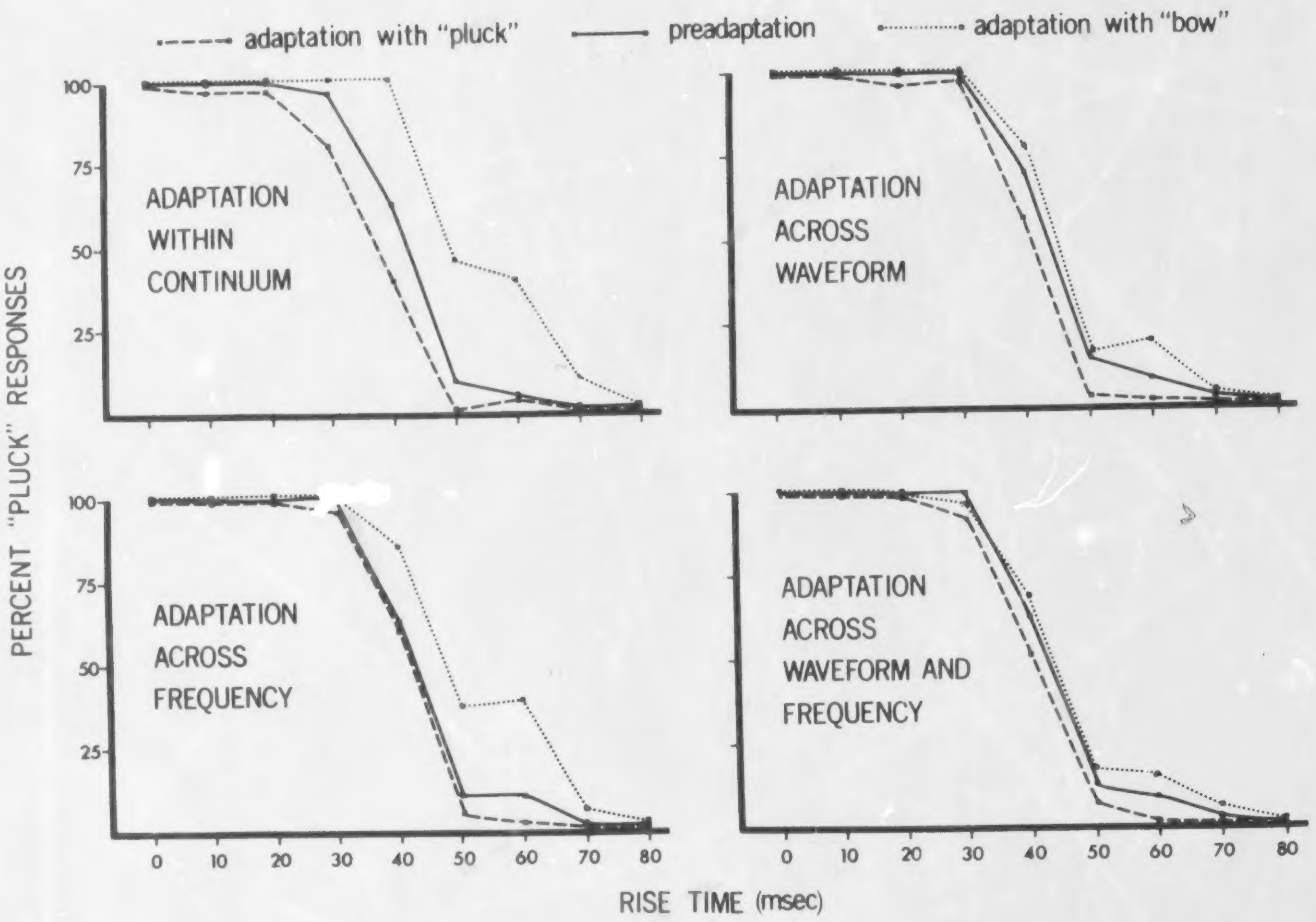


Figure 1: Mean selective adaptation functions in eight adaptation situations.

TABLE 1: Magnitudes of postadaptation boundary shifts for nonlinguistic auditory stimuli differing in rise time, in eight adaptation situations. Negative numbers indicate shifts toward 0-msec rise time, and positive numbers, shifts toward 80-msec rise time.

	Shift in Category Boundary (msec)	
Condition	Adapt at 0 msec (pluck)	Adapt at 80 msec (bow)
Adapt within continuum	-2.9 [$T(8) = 1$, $p < .02$]	10.0 [$T(8) = 0$, $p < .01$]
Adapt across frequency	.3 [$T(8) = 19$, ns]	6.6 [$T(8) = 0$, $p < .01$]
Adapt across waveform	-5.5 [$T(8) = 1$, $p < .02$]	3.5 [$T(7) = 1$, $p < .05$]
Adapt across frequency and waveform	-2.2 [$T(8) = 3$, $p < .05$]	2.2 [$T(8) = 10$, ns]

to make within-condition postadaptation functions for pluck as against bow differ significantly from one another [$T(8) = 0$, $p < .01$; $T(7) = 2$, $p < .05$; $T(8) = 1$, $p < .02$; and $T(8) = 0$, $p < .01$, for the four conditions, respectively]. In general, adaptation with a bow stimulus produced larger boundary shifts than did adaptation with a pluck stimulus. This tendency seems at least partly related to an inherent limitation in a continuum like this one: onset envelopes can be no more abrupt than 0-msec rise time but can be far more gradual than 80-msec rise time. Asymmetries have also been found in adaptation shifts with speech stimuli (Cooper, in press).

Discussion

Relatively large postadaptation shifts occur when the adapting stimulus is a member of the sawtooth array to be identified. Smaller boundary shifts appear when the adapting stimulus shares only waveform or only frequency with the test continuum, and seemingly still smaller shifts occur when that stimulus shares neither dimension. From this pattern of results we may begin to assess the abstractness of the perceptual mechanisms behind such postadaptation shifts.

The issue of abstractness is crucial. Eimas and Corbit (1973) demonstrated that postadaptation shifts could be obtained for consonant-vowel syllables arranged along a voice-onset continuum from [ba]-to-[pa] and from [da]-to-[ta]. For example, when the adapting stimulus was the most extreme token of [ba], the [ba]-[pa] phoneme boundary shifted toward the adapting stimulus. Furthermore, adaptation occurred across stimulus classes differing in place of articulation. Thus, adaptation with [da] also shifted the [ba]-to-[pa] phoneme boundary toward [ba], the labial counterpart of the adapting stimulus. Eimas and Corbit felt that such results indicated that the particular features being adapted were phonetic in nature.

Phonetic features are highly abstract (see Studdert-Kennedy, in press). The same phoneme, for example, can be manifested in entirely different acoustic forms (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). If one argues from adaptation data that *phonetic* feature detectors exist, one must demonstrate conclusively that postadaptation shifts cannot be attributed to *auditory* (acoustic) features shared between adapting and test stimuli. In this burgeoning field no demonstration seems to have sufficiently ruled out possible auditory contributions (but see Cooper, in press). Eimas and Corbit's stimuli, for example, exhibit features of voicing that are very powerful as purely auditory cues (see Miller, Pastore, Wier, Kelly, and Dooling, 1974; Stevens and Klatt, 1974).

Eimas and his colleagues have confronted this issue (Cutting and Eimas, in press) and have undertaken much research to resolve it. Eimas et al. (1973), for example, demonstrated that postadaptation shifts transfer from one ear to the other, suggesting the involvement of a single mechanism well beyond the cochlea. Thus, the locus of the effect is "abstract" enough to be removed at least several synapses from the signal. Although they obtained boundary shifts after adaptation with synthetic consonant-vowel syllables differing in voice-onset time, the effect vanished when adapting with only the initial 50 msec of the stimuli, which contained voice-onset information but sounded like chirps. Eimas et al. concluded that the perceptual mechanisms involved must be a part of the processing apparatus engaged uniquely during speech perception. Subsequently, Cooper (1974) and Ades (1974a) demonstrated that adaptation shifts also occurred across different vowel environments. Thus, for example, adaptation with [be] shifted the [bæ]-[dæ] phoneme boundary (Ades, 1974a). Such shifts, however, did not occur across syllable position. Adaptation with [bæ] had no effect on the boundary of an [æb]-[æd] continuum. The entire constellation of results suggests that postadaptation shifts are abstract enough to transfer in many but not all phonetic situations (for a review of the current literature, see Cooper, in press). A conservative conclusion, then, is that selective adaptation to speech stimuli taps abstract perceptual mechanisms that are not involved exclusively in the perception of speech.

The results of Experiment I support this view. The postadaptation shifts found for our sawtooth stimuli can hardly be interpreted as the result of phonetic feature adaptation. Moreover, such shifts in certain nonspeech continua should be expected; after all, the current work in speech adaptation stems from work in vision using nonlinguistic stimuli (McCollough, 1965; Blakemore and Campbell, 1969). Thus, the thrust of our results is twofold: (1) postadaptation boundary shifts can occur for auditory nonlinguistic stimuli just as for speech items, and (2) before concluding that any such boundary shifts for speech stimuli are explicable in linguistic terms, all possible auditory contributions must be carefully eliminated.

EXPERIMENTS II-IV: CATEGORICAL PERCEPTION

Cutting and Rosner (1974) demonstrated categorical perception for music-like stimuli differing in rise time. The results were functionally identical to those for fricatives and affricates, also cued by rise time. To probe the categorical perception of the nonlinguistic sounds in more depth, we compared them to consonants and vowels in two discrimination paradigms.

General Method

Stimuli. Two seven-item arrays of speech stimuli and one nine-item array of nonspeech stimuli were generated for identification and discrimination. One speech array consisted of speech patterns varying in direction and extent of the second- and third-formant transitions. These items were identifiable as either [bæ] or [dæ]. The second speech array consisted of three-formant steady-state vowel syllables, differing in formant frequencies and all identifiable as either [i] or [I]. The consonant-vowel continuum was generated on the Haskins Laboratories parallel-resonance synthesizer, and the vowel continuum on the vocal-tract analog synthesizer at the Research Laboratory of Electronics, Massachusetts Institute of Technology. Both continua were used previously by Pisoni (1971, 1973), who gives a more detailed description. His original stimuli were 300 msec in duration, but for the present study all items were trimmed at offset to be 250 msec in duration. The nine-item musiclike continuum from a Moog synthesizer consisted of sawtooth waves at 294 Hz differing in rise time by 10-msec increments. They were used previously by Cutting and Rosner (1974). The original stimuli were between 1020 and 1100 msec in duration, decaying in amplitude over the final second. Items here, however, were trimmed at 250 msec to conform to the duration of the speech stimuli. Thus, stimuli in all three sets had abrupt offsets. All stimuli had been digitized and stored on disc file using the pulse code modulation system at Haskins Laboratories.

Subjects and apparatus. Sixteen University of Pennsylvania undergraduates, graduate students, and secretaries were selected according to the same criteria as in Experiment I and were paid for their participation in Experiments II, III, and V. They listened in groups of four to audio tapes played on an Ampex AG500 tape recorder. Signals were sent through a listening station to matched Telephonics earphones (Model TDH-39) at 80 db re 20 $\mu N/m^2$.

Tapes and procedures. One identification tape, one variable-interval AX discrimination tape, and one ABX discrimination tape were prepared for each of the three continua. All tapes were presented diotically. Identification tapes for speech stimuli consisted of a random sequence of 70 items: (7 stimuli per array) by (10 observations per stimulus), with 3 sec between items. Subjects wrote down B or D for consonant stimuli and EE or IH for vowel stimuli. The identification tape for the sawtooth continuum consisted of 90 items: (9 stimuli) by (10 observations per stimulus), with 3 sec between items. Subjects wrote P for pluck, or B for bow after each item. After hearing five tokens of each of the full-duration endpoint items in alternating sequence, subjects readily agreed that the labels for all three types of stimuli were easy to use. They then listened to the truncated endpoint stimuli; most reported that identifiability was unimpaired.

Variable-interval AX discrimination tapes were patterned after those of Pisoni (1973). Items in the [bæ]-to-[dæ] and [i]-to-[I] arrays were numbered from 1 to 7, respectively, and the items in the pluck-to-bow array from 0 to 8. Stimuli 1, 3, 5, and 7 were then selected from each continuum. Each of these four items was paired with itself (AA pairs) and with the items adjacent to it along the abbreviated two-step continuum (in both AB and BA permutations). This process produced four AA pairs (1-1, 3-3, 5-5, and 7-7) and six AB/BA pairs (1-3, 3-1, 3-5, 5-3, 5-7, 7-5). The 3-5 and 5-3 pairs were then represented twice as often as all others (unlike Pisoni, 1973), yielding a block total of 12 pairs. The additional AB/BA pairs equalize occurrences of within- and between-category comparisons. Each trial consisted of a 100-msec 1000-Hz warning tone, followed by 750 msec of silence, followed by Stimulus A, a variable silent interval, and Stimulus X. The time interval between offset of A and onset of X was either 250, 750, or 1800 msec. Each of three AX tapes, one for each stimulus class, consisted of 72 trials in random sequence: (12 pairs per block) by (3 time delays) by (2 observations per pair), with a 4-sec interval between trials. After each trial listeners wrote *S* if they thought the items were the same, and *D* if they thought they were different.

ABX discrimination tapes were prepared with Stimuli 1 through 7 from each array. AB comparisons were selected by pairing each stimulus with the next stimulus either one or two steps removed along the continuum. Thus, there were 6 possible one-step comparisons and 5 possible two-step comparisons, for a total of 11. Each AB pair yielded four ABX arrangements: ABA, ABB, BAA, and BAB. Each tape consisted of a random sequence of 88 items: (11 comparisons) by (4 ABX permutations) by (2 observations per comparison), with 1 sec between members of triads and 4 sec between triads. Subjects wrote *A* or *B*, indicating which of the initial two items of the triad they thought identical to the third item.

Subjects performed the identification, AX, and ABX tasks in that order within one class of stimuli (consonant, vowel, or sawtooth) before listening to the next class. The order of listening to these stimulus classes and an additional one followed a Latin Square design. The fourth stimulus class and the effects it yielded are not discussed since they are not relevant to this paper. The results of the consonants and vowels are considered as Experiment II and those of the nonspeech stimuli as Experiment III.

Experiment II: Consonants and Vowels

Results and discussion. The results are remarkably similar to those of Pisoni (1973). Consonant-vowel syllables yielded categorical perception as defined by Liberman et al. (1957, 1967) and Studdert-Kennedy et al. (1970). The upper left-hand panel of Figure 2 shows the ABX discrimination results superimposed on the identification data. Note that discrimination is best at the phoneme boundary between [b] and [d]. Both one-step and two-step discrimination functions show significantly better performance at midcontinuum comparisons [$F(5, 75) = 4.61$, $p < .01$,and $F(4, 60) = 19.3$, $p < .001$, respectively]. The results of the variable-interval AX task are shown in the lower left-hand panel of Figure 2. As in the ABX task, the 3-5 comparisons were much easier than the 1-3 and 5-7 comparisons [$F(1, 15) = 75.2$, $p < .001$]. The duration of the silent interval between items A and B did not significantly influence judgments.

The complementary vowel data appear at the right-hand side of the same figure. They are less categorical. The discrimination data are superimposed on the identification fucntions; again, both one- and two-step discrimination functions show significant peaks at the vowel boundary [$F(5, 75) = 3.71$, $p < .01$, and $F(4, 60) = 6.57$, $p < .01$, respectively]. Discrimination performance, however, is markedly better for vowels than for consonants [$F(1, 15) = 38.2$, $p < .001$, and $F(1, 15) = 39.1$, $p < .001$, for one- and two-step comparisons]. The one-step vowel and one-step consonant functions do not differ significantly in shape, since there is no Stimulus Class by ABX Comparison Interaction. The two-step vowel function, however, is markedly "flatter" than the two-step consonant function [$F(4, 60) = 6.84$, $p < .001$]. This result emphasizes the importance of one criterion for categorical perception: the presence of troughs in the discrimination functions where within-category comparisons yield chance performance. Such results occur consistently for consonant but not for vowel discriminations.

The variable-interval AX results reveal a different time course for perception of vowels as against consonants. Both vowels and consonants demonstated

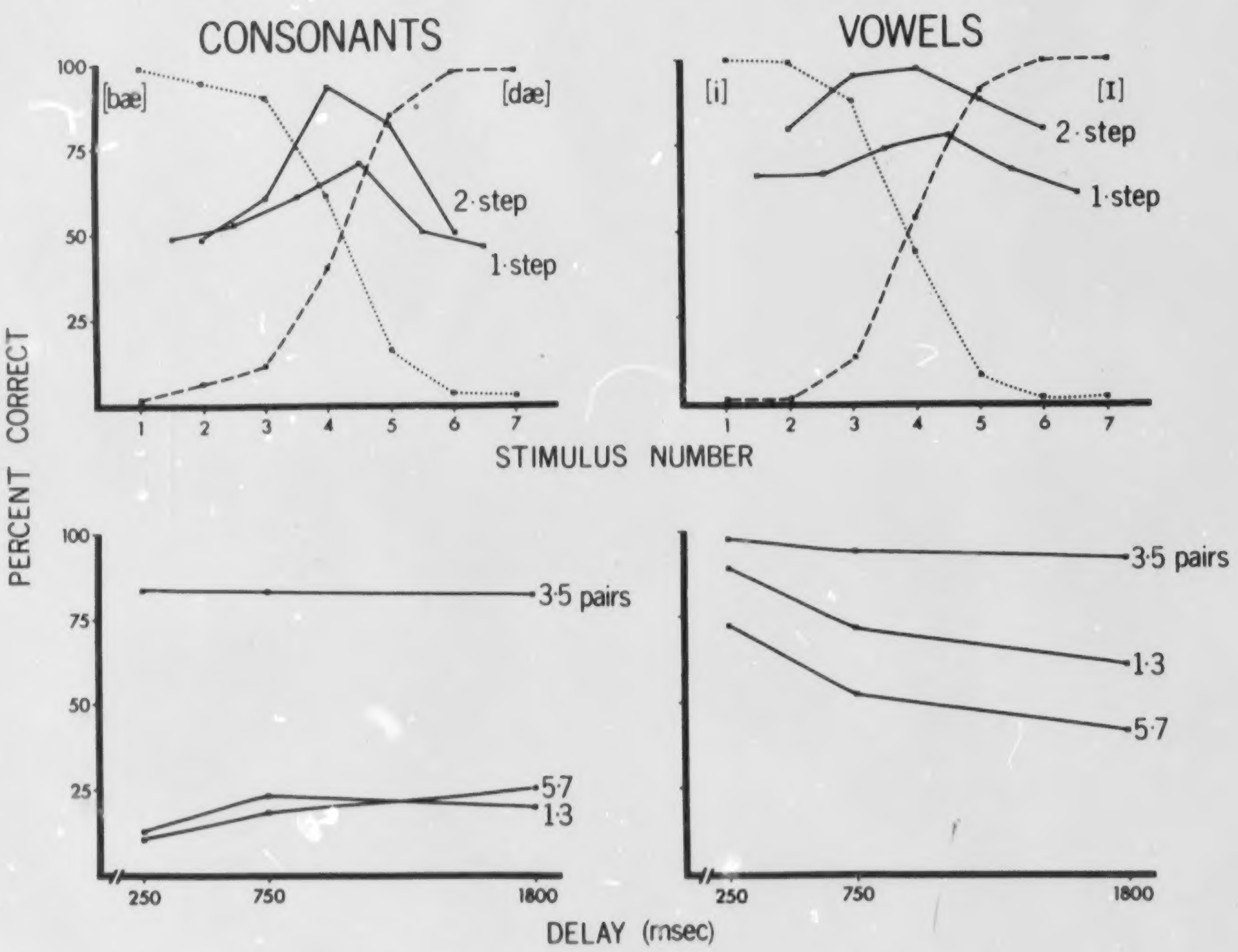


Figure 2: Mean identification, ABX discrimination, and variable interval AX discrimination functions for consonants and vowels.

differential difficulty among the three comparisons: the 3-5 comparison was considerably easier to judge than the 1-3 and 5-7 comparisons taken together [$F(1, 15) = 52.3$, $p < .001$]. However, unlike the consonants, the vowels demonstrate a significant difference in discriminability as a function of silent interval between A and B items [$F(2, 30) = 8.25$, $p < .01$]. Furthermore, the Stimulus Class by AB Comparison by Delay Interval Interaction was significant [$F(2, 30) = 12.9$, $p < .001$]. This finding suggests that within-category acoustic information decays for vowels but not for consonants. With consonants the differential information is "lost" prior to the onset of the second item in the AX pair. Our results for consonants and vowels are very similar to those of Pisoni (1973). He also plotted AX results in terms of d'. Such a plot of our data yields no patterns besides those already apparent in the lower panels of Figure 2.

Experiment III: 250-msec Sawtooth Stimuli

Since the [bæ]-to-[dæ] stimuli yielded results indicative of categorical perception, while the [i]-to-[I] stimuli yielded a less categorical outcome, these data provide a yardstick for assessing categorical perception of nonspeech pluck-to-bow stimuli varying in rise time. Previously, Cutting and Rosner (1974) found that in an ABX task these items yielded categorical perceptions nearly identical to those of affricate-vowel and fricative-vowel syllables also differing in rise time. However, two important considerations suggest that these musiclike stimuli might not have produced categorical perceptions functionally identical to those for consonants.

First, consider the previous results for a sawtooth wave continuum whose average item duration was 1060 msec; the data appear in the upper left-hand panel of Figure 3. The two-step discrimination function is overlaid on the identification function just as before. Since item duration slightly exceeded 1 sec and since intervals between items in ABX triads were 1 sec, one might argue that the within-category troughs of these data reflected the 2-sec interval between onsets of items A and B. Our AX vowel discrimination data and those of Pisoni (1973) indicate that an onset-to-onset delay of 2 sec could decrease within-category discriminability to near chance, even though strongly categorical perception is absent.

Second, Cutting and Rosner's control stimuli were affricates and fricatives. Liberman et al. (1967) noted that such speech segments are "less encoded" than stop consonants. Darwin (1971) demonstrated in a dichotic listening task that fricatives can yield small ear advantages more similar to those found for vowels than for stop consonants. Comparing results for the sawtooth stimuli to those for fricatives and affricatives may be too weak a test of categorical perception for the musical sounds.

Thus, in the present study we shortened the sawtooth stimuli to 250 msec, the same duration as the speech stimuli in Experiment II. In addition to obtaining identification and discrimination functions, we intended to observe whether the time course of within-category discriminations followed that for the more categorical stop consonants or for the less categorical vowels.

Results and discussion. The initial outcome was discouraging, as the right-hand side of Figure 3 shows. The identification results were not nearly

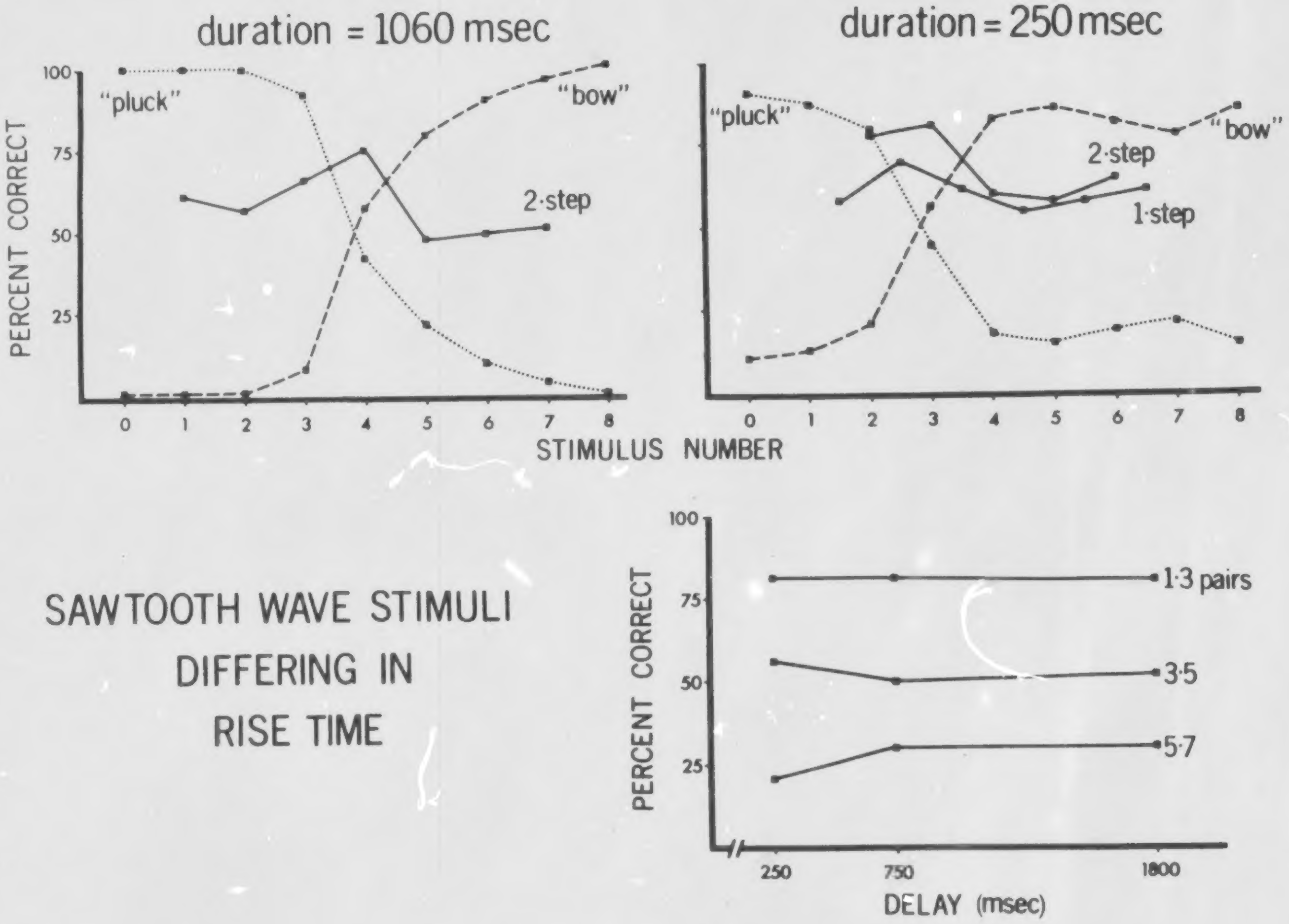


Figure 3: Mean identification and ABX discrimination functions for the full-duration sawtooth stimuli (Cutting and Rosner, 1974) compared against the mean identification, ABX and AX discrimination functions for the 250-msec sawtooth stimuli.

as quantal as those of Cutting and Rosner, which appear on the left-hand side. Moreover, the boundary for the 250-msec stimuli, as indicated by the complementary pluck and bow identification functions, appears to be around 30-msec rise time, whereas the 1060-msec stimuli produced a boundary at about 40 msec. Other differences characterize the ABX discrimination results. The one-step discrimination function was nearly flat with at best only a slight peak at the 2-3 comparison. The two-step function revealed some perturbations that might indicate a boundary at about 30-msec rise time. Nevertheless, no adequate within-category comparison is available at short rise times, a result of the apparent boundary relocation. Thus, the data cannot be said to show categorical perception. The AX discrimination data display the anticipated lack of decay of within-category information, but the notion of perceptual categories for these 250-msec items still may be wrong.

One reason for this conservative assesssent is that Figure 3 shows group data for all 16 subjects. Despite their claims during preliminary familiarization with the stimuli, six listeners could not perform the main experimental task of systematically identifying the sounds as pluck and bow. This added considerable noise to the data. Moreover, these listeners performed at chance in both ABX and AX discrimination tasks. This large minority of subjects unable to do the tasks forces us to conclude that no readily apparent categories exist for these stimuli: the perception of the truncated items diverges markedly from the perception of the original stimuli.

Thus, we held the results for consonants and vowels in abeyance while we investigated the reason for the discrepancy between the results for 250-msec sawtooth stimuli and the prior findings for the same items at durations exceeding 1 sec.

Experiment IV: 750-msec Sawtooth Stimuli

Method. The original Cutting and Rosner sawtooth items were truncated again, but this time at 750 msec rather than 250 msec. Identification, ABX and AX discrimination tapes were then recorded using the same test orders as for the 250-msec items. Eight of the 16 subjects in Experiments II and III were recalled and paid to perform the same three tasks with the 750-msec items as they had previously with the shorter items. Among the eight listeners were four who did not consistently perform the tasks in Experiment III. Each subject listened individually using the same apparatus as in Experiment I.

Results. The results for the 750-msec sawtooth items are shown at the left-hand side of Figure 4. They can be compared against those for the 250-msec items for the same eight listeners shown at the right-hand side. The identifications of the 750-msec stimuli are considerably more quantal than those of the 250-msec items. Moreover, the boundary designated by the complementary identification functions has moved back to the 35-to-40-msec rise time found by Cutting and Rosner (1974). This change from the shorter-item boundary was significant [$F(8, 56) = 3.68$, $p < .01$].

The one- and two-step ABX functions also indicate a more categorical type of perception. Both show significant midcontinuum peaks [$F(5, 35) = 4.94$, $p < .01$, and $F(4, 28) = 10.26$, $p < .001$, respectively], and both are significantly different from those for the 250-msec items [$F(5, 35) = 2.66$, $p < .05$, and $F(4, 28) = 4.83$, $p < .01$).

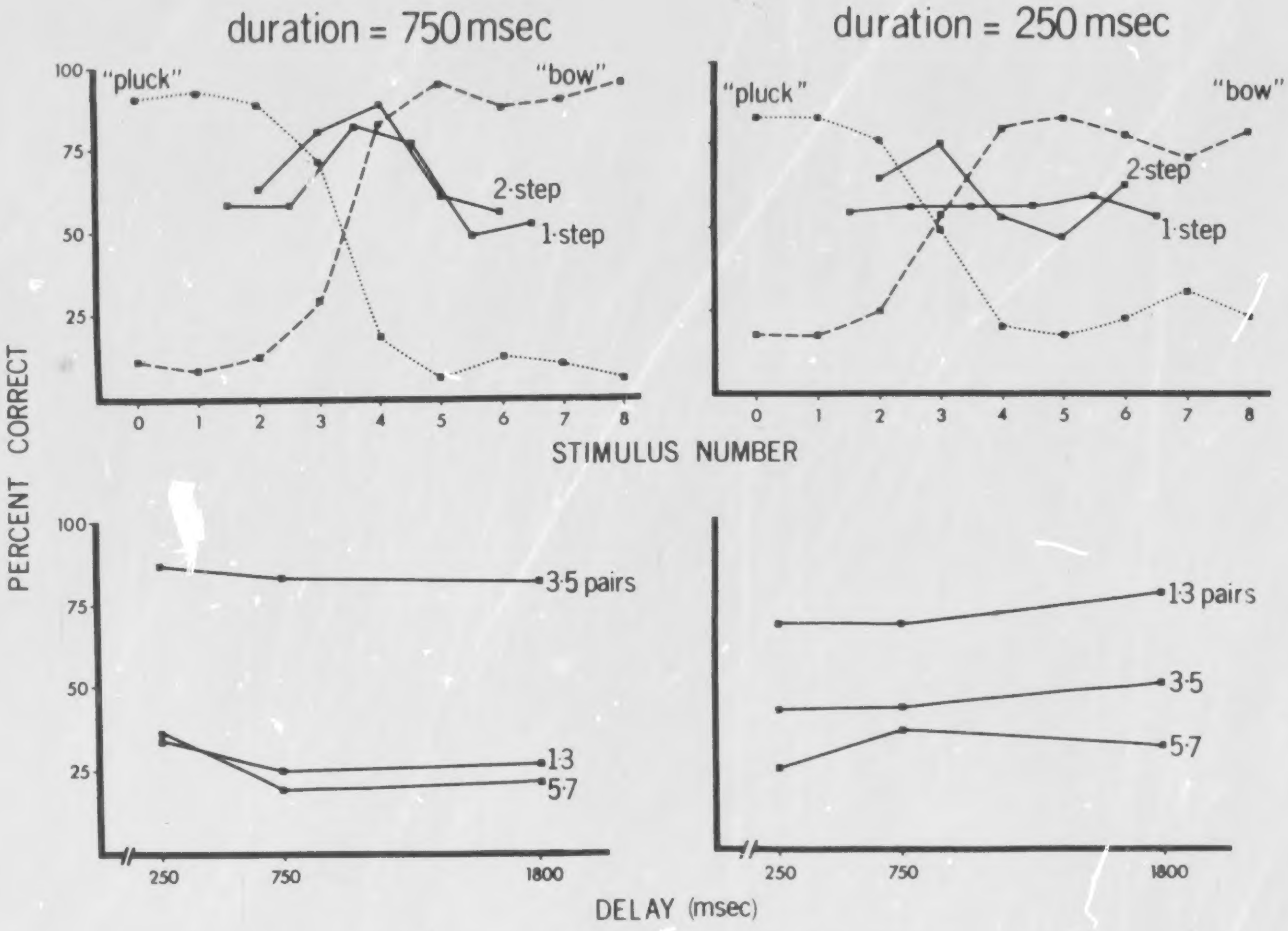


Figure 4: Mean identification, ABX and AX discrimination functions for the sawtooth items truncated at 750 and 250 msec for 8 of the 16 subjects contributing to the data shown at the right-hand side of Figure 3.

The AX discriminations follow suit. Whereas the 250-msec 1-3 pairs were discriminated most easily in Experiment II, the 750-msec 3-5 pairs were much easier to discriminate than the other two pairs in this experiment. This change in arrangement of pair discriminability was highly significant [$F(2, 14) = 11.78$, $p < .001$]. The 750-msec items follow a pattern just like that for the stop consonants shown in Figure 2. Unlike the consonants, however, the 750-msec sawtooth items did show some slight decay of information over time [$F(2, 14) = 3.44$, $p < .05$], behaving slightly like vowels. For the eight subjects in this experiment, however, the sawtooth items do not significantly differ in such decay from the consonant stimuli. Thus, the principal outcome is that sufficiently long-lasting rise time items are perceived and categorized in a manner considerably more like that of stop consonants than that of vowels.

Discussion. We first note an interesting discrepancy between these results for musiclike sounds and Pisoni's (1973) findings with synthetic speech. In his experiments, the perception of short vowels (50 msec) was more categorical than that of longer vowels (300 msec). For pluck and bow stimuli we found the opposite result: the longer stimuli (750 msec) yielded far more categorical-like perceptions than the shorter items (250 msec).

Pisoni accounted for his results between vowels of different durations in terms of the amount of information available in short-term memory. Because the 50-msec vowels are shorter in duration, they must be encoded more rapidly from the information in short-term store. The rapid encoding process appears to contribute to categorical perception. Abbreviating the stimuli from 300 to 50 msec, however, did not impair vowel identifiability. In contrast, abbreviation from 1060 (or even 750) msec to 250 msec did impair identifiability and discriminability of our sawtooth stimuli. We have no fully adequate explanation at present for this effect. There seem to be at least two approaches here. First, trimming the stimuli to 250 msec eliminates the timbre associated with a stringed instrument. Indeed, many subjects reported that the short items sounded more like quick toots on a harmonica. Second, the abrupt offset of each item is quite disruptive and may even "mask" critical information about onset rise time. Either of these explanations might account for the fact that identifications of the 250-msec items were less quantal than those of the 750-msec items, which in turn were somewhat less quantal than those of the full 1060-msec items of Cutting and Rosner (1974) and of Experiment I. In any event, our results emphasize that onset cues to the timbre of musical sounds require subsequent acoustical events in order to be effective. The shift in boundary between categories as stimulus duration increases clearly manifests this effect. Parallel interactions occur in the perception of speech (Mattingly, Liberman, Syrdal, and Halwes, 1971), where, for example, formant transition cues for stop consonants are heard as chirps by themselves. These cues require subsequent steady-state vocalic formants (as well as first-formant transitions) in order to be heard as speech sounds.

The results of Experiment IV strongly support the conclusion of Cutting and Rosner: categorical perceptions of nonspeech stimuli differing in rise time are functionally identical to those of certain speech sounds. In fact, they are nearly identical to those of the most categorical of speech sounds--stop consonants, as shown in Experiment II--in that there is only minimal decay of within- and between-category information over the time intervals 250 to 1800 msec. Still, one might argue that the results of Experiment IV are inconclusive.

Because sawtooth items required lengthening to 750 msec, within-category decay of information, even at 250-msec offset-to-onset delays, may have already reached asymptote. Since the stimuli differ in onset and last 750 msec, critical comparisons at nominal 250-msec delays are really made between onsets separated by 1.0 sec. Likewise, critical comparisons at 750- and 1800-msec delays are really made between onsets separated by 1.5 and 2.55 sec. The perception of stimuli differing in rise time thus could be more similar to that of vowels than consonants. Discrimination performances for onsets separated by 1.0, 1.5, and 2.55 sec may asymptote in less than 1 sec resulting from decay of crucial within-category information.

There are two arguments against this possibility. First, if within-category information were to decay for the sawtooth items as it does for vowels, then Pisoni's (1973) data clearly call for performance differences even at such long onset-to-onset delays. Second, the discrimination performance for within-category sawtooth pairs at a 250-msec offset-to-onset delay is already below the within-category performance for vowels at a 1800-msec offset-to-onset delay. (Compare Figures 2 and 4.)

Because data on selective adaptation and on categorical perception with the sawtooth stimuli were so compelling, we looked for perceptual similarities in dichotic listening between the stop consonants and pluck and bow stimuli.

EXPERIMENTS V AND VI: DICHOTIC LISTENING

General Method

Two dichotic tapes were prepared, one consisting of natural speech tokens of [ba, da, ga, pa, ta, ka], each between 270 and 300 msec in duration, and the other consisting of the pluck and bow musiclike items lasting about 1 sec. The speech tape consisted of simultaneous-onset pairs in which every item was combined with each other item except for itself. A random sequence of 60 pairs was recorded using the PCM system at Haskins Laboratories: (15 possible dichotic pairs) by (2 channel arrangements) by (2 observations per pair). The nonspeech tape consisted of the eight-endpoint full-duration tokens of the rise time stimuli used by Cutting and Rosner: items with 0- and 80-msec rise times at 294 and 440 Hz from both sawtooth and sine wave continua. Like the speech tape, every item was paired with each other item except for itself. A random sequence of 112 simultaneous-onset dichotic pairs was recorded: (28 possible pairs) by (2 channel arrangements) by (2 observations per pair).

Experiment V: Ear Differences

The same 16 subjects who participated in Experiments II and III carried out two tasks here. They monitored a given ear for a particular block of trials and wrote down the item they heard: initial B, D, G, P, T, or K, for speech task, and P (pluck) or B (bow) for the nonspeech task. Subjects listened to each tape twice, reversing headphones after the first pass through the tape. Half the subjects monitored the right ear for the first quarter of the task, then the left ear for the next two quarters, and then the right ear again for the final quarter: RLLR. The other subjects monitored in the opposite order, LRRL. Headphone configurations and the order in which subjects did the tasks were counterbalanced across subjects. The apparatus for Experiment II was used here as well. Items were again presented at 80 db re 20 $\mu N/m^2$.

Results and discussion. The consonant-vowel syllables yielded a right-ear advantage, but the pluck and bow musiclike sounds did not. For speech stimuli, subjects were 76 percent correct when monitoring the right ear and only 66 percent correct when monitoring the left, a net 10 percent right-ear advantage. Twelve of 16 subjects yielded results in this direction [$z = 1.87$, $p < .06$, by a two-tailed sign test) and the mean ear advantage would have been much greater had not one subject yielded a very large left-ear advantage ($z = -4.38$, $p < .0001$). Nonspeech stimuli, on the other hand, yielded no ear advantage: listeners were 70 percent correct when monitoring both left and right ears.

Individual subject scores were then converted to phi coefficients (Kuhn, 1973) and speech and nonspeech results were compared. Eleven of 15 subjects yielded results that were suggestive of larger right-ear advantages for speech than for nonspeech ($z = 1.68$, $p < .10$). (For the remaining subject the phi coefficients were the same.) Since this Ear Difference by Stimulus Class Interaction was suggestive but not significant at the .05 level, we planned a final study to assess the reliability of ear advantages for both speech and sawtooth stimuli.

Experiment VI: Reliability of Ear Differences

The tapes for the previous experiment were used here. Eight of the 16 subjects were recalled and paid to repeat Experiment V. The apparatus was the same as in Experiment I and subjects were tested individually. Procedures were otherwise identical to those of Experiment V.

Results and discussion. Ear scores of the eight subjects for speech and nonspeech stimuli in both Experiments V and VI were converted into z scores and appear in Table 2. A Spearman rank-order correlation revealed that the ear differences for the speech stimuli were quite reliable [$r_s(8) = .86$, $p < .01$], whereas the ear differences for the nonspeech stimuli were not [$r_s(8) = .45$, ns], suggesting only regression toward a mean value of 71 percent correct for both ears.

TABLE 2: Ear advantages in terms of z scores for speech stimuli and sawtooth stimuli, as an assessment of test-retest reliability. Negative scores indicate left-ear advantages and positive scores, right-ear advantages.

	z Score			
	Speech Stimuli		Sawtooth Stimuli	
Subject	Experiment V	Experiment VI	Experiment V	Experiment VI
SM	3.49	6.35	- .59	1.11
DE	3.24	5.34	1.56	- .15
SE	2.75	.21	.59	.83
MM	2.12	4.48	1.52	.00
MMc	1.37	3.40	1.34	.00
GM	1.07	.39	-1.48	-1.91
LT	-1.49	- .19	-1.32	- .62
RM	-4.38	-2.66	-2.49	-2.08

The results of Experiments V and VI, then, demonstrate that our musiclike items do not yield speechlike results in all situations. They yielded neither a right-ear advantage, a typical result for speech stimuli (Kimura, 1961; Chaney and Webster, 1965; Studdert-Kennedy and Shankweiler, 1970), nor a left-ear advantage, a result reported in some cases for nonspeech stimuli (Kimura, 1964; Chaney and Webster, 1965; Gordon, 1970). We have no evidence that unique cortical processing in either hemisphere occurs for our stimuli. The negative results of these two dichotic studies require careful interpretation. There were only two possible responses in the nonspeech task, pluck or bow, as against six in the speech task. A two-item repertory of responses might reduce the size of any real ear advantage; the subject has a less demanding judgment to make. Nevertheless, experiments with two-choice responses have yielded ear advantages. Chaney and Webster (1965) found a significant right-ear advantage with just the vowels [i] and [a] and a left-ear advantage with just a sonar reverberation and a cry of a humpbacked whale. Moreover, Cutting (1974b, Experiment II) employed an ear-monitoring task identical to that in the present studies in which the only possible responses were [b] or [g], and yet he found a significant 6 percent right-ear advantage. (This is somewhat less than our Experiment V with speech stimuli.) We infer that our results with musical stimuli indicate less highly lateralized cortical processing than for speech, and perhaps no lateralization at all.

The present study in conjunction with Experiments II and IV supports the conclusion of Cutting (1974a). He demonstrated that categorical perception and the right-ear advantage did not necessarily appear together. Right-ear advantages occurred for both consonant-vowel syllables and pseudo-syllables resembling them but with inverted first-formant transitions. Only the consonant-vowel syllables, however, yielded categorical perception. Our studies produced exactly the opposite pattern: consonant-vowel syllables and pluck and bow sawtooth items yielded categorical perception, but only the speech items yielded a clear right-ear advantage.

A LOOK AT SOME MODELS OF SPEECH PERCEPTION

If man's productive apparatus sets his speech capacities apart from those of other animals (Lieberman, 1973), does his perceptual apparatus have unique advantages as well? Some 20 years of research have produced several experimental findings that conceivably point in this direction. Three are directly relevant here: (1) boundary shifts in selective adaptation for certain speech continua, (2) categorical perception as reflected by identification and discrimination functions, and (3) the right-ear advantage in dichotic listening. Let us consider each in some detail, with specific attention to models offered to account for them.

Selective Adaptation

A category boundary that shifts after selective adaptation obviously presupposes the existence of discrete categories. Thus, to demonstrate such shifts among nonspeech sound, even approximately like those found in speech, one must discover a nonlinguistic but categorically perceived auditory dimension. Such dimensions are precious few in number at the present time, but rise time is clearly one of them.

Eimas and Corbit (1973) did not specify that postadaptation shifts must be phonetic in nature; they merely interpreted those that they found to be phonetic. The results of Experiment I suggest that boundary shifts occur readily for nonlinguistic stimuli. The mechanisms underlying shifts for certain speech stimuli, then, may be auditory just as they must be auditory for pluck and bow sounds. Such a conservative conclusion seems reasonable given the fact that these adaptation effects were first noted in vision for stimuli carrying no linguistic information.

Nevertheless, theories of speech perception based on results of selective adaptation can accommodate our findings. Ade (1974b), for example, has proposed a two-tiered feature detection model: one tier of detectors is auditory in nature and the other phonetic. This is a logical improvement over the notion of purely phonetic-feature detection since there are many acoustic manifestations of a particular phoneme. Perhaps the outputs of different sets of auditory detectors are directed into a more abstract phoneme detector according to phonological context.

The results of Experiment I, however, suggest that a single tier of auditory detectors is insufficient. Adaptation shifts are greater when the adapting stimulus shares all dimensions with the test continuum than when only frequency, or only waveform, or neither is shared. A model accounting for these results would appear to need at least two auditory tiers or levels. The first might handle such features as pitch or waveform, as well as onset envelope, and all of these features might map onto a second tier where a binary decision is made. The more lower-tier features that are shared between adapting and test stimuli, the larger the adaptation effect at the second tier. We cannot formulate a complete model to account for our results, much less all of those found with speech syllables. It is clear, however, that tiers of auditory property detectors must proliferate. With this proliferation, a feature detection model of speech perception will quickly become complex, somewhat cumbersome, and less appealing.

If such multitiered devices can account for categorization processes in speech perception, and in some sense they must, the perception of speech would surely use all available auditory detectors along with some that may be unique to speech. Thus, it would not be the use but the extended use of such devices that is unique to speech. However, there is now a pressing need to verify the existence of phonetic feature detectors as distinct from phonetically relevant auditory detectors.

Categorical Perception

Categorical perception is clearly not unique to speech. The results of Experiment II-IV in the present paper, and those of Cutting and Rosner (in press), indicate that sawtooth stimuli differing in rise time are perceived categorically according to the strictest criteria (Studdert-Kennedy et al., 1970; Pisoni, 1973). Our nonspeech stimuli, however, are not alone in this regard. Miller et al. (1974) varied the onset timing of a hiss and a buzz, emulating the aperiodic and periodic aspects of consonants along a voice-onset continuum. They found that these sounds also were categorically perceived. To a somewhat lesser degree, Locke and Kellar (1973) uncovered categorical perceptions in trained musicians who listened to musical triads differing in their middle component. Finally, Lane (1965) reported near-categorical perceptions of inverted speech patterns by pretrained listeners (but see Studdert-Kennedy et al., 1970).

Nevertheless, categorical perception *seems* more characteristic of speech sounds than of nonspeech sounds. The most complete model of the process stems from the work of Fujisaki and Kawashima (1968, 1970) as amplified by Pisoni (1971, 1973, in press). The peaks and troughs in the ABX discrimination functions of continua such as [i]-to-[I] or [bæ]-to-[dæ] appear to result from relative strengths of information in different memory stores. The depth of the troughs (regions of discriminability that remain marginally above chance) partly represent the strength of auditory memory. The deeper the troughs, the less the within-category stimulus information has remained in a relatively raw acoustic form. Thus, as Pisoni (1973) has shown, long vowels are less categorical than short vowels because of differential auditory information. For the same reason, short vowels are less categorical than stop consonants. Our results for pluck and bow musiclike items fit this part of the model.

A problem arises, however, with the theoretical mechanism for the peaks in the ABX discrimination function. Peaks occur at category boundaries. For speech stimuli each boundary is *phonetic*, and thus a phonetic memory store lasting longer than an auditory store was thought to account for the higher performance here than within a category. The results of Experiments II-IV suggest that this memory need not be phonetic but could be a storage system reserved for highly coded information. The binary choice of pluck versus bow might qualify as a highly coded perceptual decision. With this modification, the Fujisaki and Kawashima model will explain the observed phenomena.

A more global view of categorical perception notes the intimate relationship between perception and production (Liberman et al., 1967): we may perceive a [bæ]-to-[dæ] acoustic continuum in a discrete manner partly because we cannot easily produce any sound intermediate between [b] and [d]. The view has great intuitive appeal with regard to speech but becomes inappropriate when extended to our pluck and bow sounds. One must argue that difficulties in producing a note on a violin between a pluck and a bow have accordingly influenced perceptions. Certainly man did not evolve with a stringed instrument tucked under his chin. Our findings offer small comfort to anyone holding a teleological theory of the relationship of perception and production.

Categorical perception is a manifestation of a much broader process in speech: the segmentation of a continuous auditory stream into discrete phonetic elements (Studdert-Kennedy, in press). In perceiving the initial phoneme in the syllable [bæ], we make not one but at least three binary decisions. The consonant is [b] not its voiceless counterpart [p], nor its alveolar neighbor [d], nor its nasal relative [m]. The first two distinctions are clearly categorical (Pisoni, 1971, 1973), and the third may be. Three binary categorical decisions made in the processing of a syllable that can easily be uttered in 100 msec would seem to indicate a 30 bit-per-second process. Yet this bit rate may be an *under*estimate for that of running speech; Liberman, Mattingly, and Turvey (1972), for example, have estimated that the coding of continuous speech can reach 40 bits per second. Since our sawtooth items truncated at 250 msec are not highly identifiable as pluck or bow, one might infer that the coding of such simple musiclike sounds does not even reach 4 bits per second, an order of magnitude less than that for speech. Adding other decisions about pitch, duration, and harmony would certainly raise this rate. But even with such additions, our ability to discretely categorize musical and other nonspeech stimuli may not approach that for speech. More broadly, then, perhaps it is not the existence but the extent of categorical perception that is unique to speech. This possibility needs further study; at present, consonants and certain nonspeech sounds do not appear to differ with regard to categorical perception.

Right-Ear Advantages and Lateralization

Although we have presented no counterevidence in the present paper, the right-ear advantage also does not appear to be unique to speech. Halperin, Nachshon, and Carmon (1973), for example, found right-ear advantages for complex noiselike patterns; Bever and Chiarello (1974) found them in musicians for the perception of melodies and melodic excerpts; and Cutting (1974b) reported a component of the right-ear advantage attributable to rapid frequency modulation.

Differential ear scores imply lateralized underlying neural processes (Kimura, 1961, 1967; Studdert-Kennedy and Shankweiler, 1970). Neurological observations (Penfield and Rasmussen, 1949; Wada and Rasmussen, 1960) and electrophysiological evidence (Wood, Goff, and Day, 1971), along with that from dichotic listening, strongly support the notion that the left hemisphere is specialized for speech. Speech and language, however, are not the only lateralized functions in man. Semmes (1968) has argued that "focal" processes are the province of the left hemisphere and that "diffuse" processes are more the property of the right hemisphere. Bever and Chiarello (1974) termed this dichotomy "analytic" versus "holistic." Such a view, although perhaps oversimplistic, would predict right-ear advantages in dichotic tasks for nonlinguistic sounds requiring more than global processing.

In Experiments V and VI, however, we found no ear advantages for the perception of nonspeech sounds differing in rise time. We recognize the problems involved generally in asserting the null hypothesis and particularly in the possible effect of response-set size on ear advantages. Nevertheless, the binary perceptual decision of pluck versus bow apparently involves little or even no lateralized cortical processing. Instead, it would seem to need only the most rudimentary analysis. Perhaps the underlying mechanisms are part of a system phylogenetically older than that which evolved to perform "analytic" or "focal" processing. This older system may be connected with orienting actions.

Conclusion

The results of the present studies speak for dissociating particular models of categorical perception and of selective adaptation from general models of hemispheric functioning. Nonlinguistic stimuli differing in rise time exhibit both categorical perception and boundary shifts associated with adaptation but do not exhibit strong lateralization in dichotic listening. Thus, the particular mechanisms involved in categorization and in shifting boundaries may not be unique to speech processing but may be part of the more general auditory processing system. Speech processing may call on these mechanisms to a greater degree or at a higher neural level than the processing of other sounds. As yet there are not enough data to develop this view properly.

Stimuli identifiable as pluck and bow are functionally identical to stop consonants in discrimination and adaptation paradigms, and musical and other nonlinguistic stimuli can yield results in dichotic listening identical to those of speech sounds. There may be no results and no mechanisms that are unique to speech perception. Lieberman (1973; Lieberman et al., 1972) has argued that it is the configuration of the human vocal tract, but not the existence of specific anatomical devices within it, that is unique to man and that enables him to speak. Analogously, perhaps it is the configuration of perceptual mechanisms but not the particular devices themselves that enables man to comprehend his own rapid speech.

REFERENCES

Ades, A. E. (1974a) How phonetic is selective adaptation? Experiments on syllable position and vowel environment. Percept. Psychophys. 16, 61-66.

Ades, A. E. (1974b) Bilateral component in speech perception? J. Acoust. Soc. Amer. 56, 610-616.

Bever, T. G. and R. J. Chiarello. (1974) Cerebral dominance in musicians and nonmusicians. Science 185, 537-539.

Blakemore, C. and F. W. Campbell. (1969) On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images. J. Physiol. 203, 237-260.

Chaney, R. B. and J. C. Webster. (1965) Information in certain multidimensional sounds. J. Acoust. Soc. Amer. 40, 447-455.

Cooper, F. S. and I. G. Mattingly. (1969) A computer-controlled PCM system for the investigation of dichotic speech perception. J. Acoust. Soc. Amer. 46, 115(A).

Cooper, W. E. (1974) Adaptation of phonetic feature analyzers for place of articulation. J. Acoust. Soc. Amer. 56, 617-627.

Cooper, W. E. (in press) Selective adaptation to speech. In Cognitive Theory, ed. by F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindman, and D. B. Pisoni. (Potomac, Md.: Erlbaum Associates), vol. 1.

Cutting, J. E. (1974a) Different speech processing mechanisms can be reflected in discrimination and dichotic listening tasks. Brain Lang. 1, 363-373.

Cutting, J. E. (1974b) Two left-hemisphere mechanisms in speech perception. Percept. Psychophys. 16, 601-612.

Cutting, J. E. and P. D. Eimas. (in press) Phonetic feature analyzers and the processing of speech by infants. In The Role of Speech in Language, ed. by J. F. Kavanagh and J. E. Cutting. (Cambridge, Mass.: MIT Press).

Cutting, J. E. and B. S. Rosner. (1974) Categories and boundaries in speech and music. Percept. Psychophys. 16, 564-570.

Darwin, C. J. (1971) Ear differences in the recall of fricatives and vowels. Quart. J. Exp. Psychol. 23, 46-62.

Eimas, P. D., W. E. Cooper, and J. D. Corbit. (1973) Some properties of linguistic feature detectors. Percept. Psychophys. 13, 247-252.

Eimas, P. D. and J. E. Corbit. (1973) Selective adaptation of linguistic faature detectors. Cog. Psychol. 4, 99-109.

Fouts, R. S. (1973) Acquisition and testing of gestural signs in four young chimpanzees. Science 180, 978-980.

Fujisaki, H. and T. Kawashima. (1968) The influence of various factors on the identification and discrimination of synthetic speech sounds. Reports of the Sixth International Congress on Acoustics (University of Tokyo), pp. 67-73.

Fujisaki, H. and T. Kawashima. (1970) Some experiments on speech perception and a model for the perceptual mechanisms. Annual Report of the Engineering Research Institute (University of Tokyo) 29, 207-214.

Gardiner, R. A. and B. T. Gardiner. (1969) Teaching sign language to a chimpanzee. Science 165, 664-672.

Gordon, H. W. (1970) Hemispheric asymmetries in the perception of musical chords. Cortex *6*, 387-398.

Halperin, Y., I. Nachshon, and A. Carmon. (1973) Shift in ear superiority in dichotic listening to temporal pattern nonverbal stimuli. J. Acoust. Soc. Amer. *53*, 46-50.

Kimura, D. (1961) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. *15*, 166-171.

Kimura, D. (1964) Left-right differences in the perception of melodies. Quart. J. Exp. Psychol. *16*, 355-358.

Kimura, D. (1967) Dual functional asymmetry of brain in visual perception. Neuropsychologia *4*, 275-285.

Kuhn, G. M. (1973) The phi coefficient as an index of ear differences in dichotic listening. Cortex *9*, 447-457.

Lane, H. (1965) Motor theory of speech perception: A critical review. Psychol. Rev. *72*, 275-309.

Liberman, A. M., F. S. Cooper, D. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. *74*, 431-461.

Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. *54*, 358-368.

Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In *Coding Processes in Human Memory*, ed. by A. W. Melton and E. Martin. (Washington, D.C.: V. H. Winston & Sons), pp. 307-334.

Lieberman, P. (1973) On the evolution of language: A unified view. *Cognition* *2*, 59-94.

Lieberman, P., E. S. Crelin, and D. H. Klatt. (1972) Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. Amer. Anthropol. *74*, 287-307.

Locke, S. and L. Kellar. (1973) Categorical perception in a nonlinguistic mode. Cortex *9*, 355-369.

Mattingly, I. G., A. M. Liberman, A. K. Syrdal, and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. *2*, 131-157.

McCollough, C. (1965) Color adaptation of edge-detectors in the human visual system. Science *149*, 1115-1116.

Miller, J. D., R. E. Pastore, C. C. Wier, W. M. Kelly, and R. M. Dooling. (1974) Discrimination and labeling of noise-buzz sequences with varying noise-lead times. J. Acoust. Soc. Amer. *55*, 390(A).

Penfield, W. and T. Rasmussen. (1949) Vocalization and arrest in speech. Arch. Neurol. Psychiat. *61*, 21-27.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Ph.D. dissertation, University of Michigan (Psycholinguistics). (Issued as Supplement to Haskins Laboratories Status Report on Speech Research.)

Pisoni, D. B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. *13*, 253-260.

Pisoni, D. B. (in press) Auditory short-term memory and vowel perception. Mem. Cog.

Premack, D. (1971) Language in chimpanzee? Science *172*, 808-822.

Semmes, J. (1968) Hemispheric specialization: A possible clue to mechanism. Neuropsychologia *5*, 11-26.

Stevens, K. N. and D. H. Klatt. (1974) Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Amer. *55*, 653-659.

Studdert-Kennedy, M. (in press) Speech perception. In *Contemporary Issues in Experimental Phonetics*, ed. by N. J. Lass. (Springfield, Ill.: C. C Thomas).

Studdert-Kennedy, M., A. M. Liberman, K. S. Harris, and F. S. Cooper (1970) Motor theory of speech perception: A reply to Lane's critical review. Psychol. Rev. 77, 234-249.

Studdert-Kennedy, M. and D. P. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.

Wada, J. and T. Rasmussen. (1960) Intracarotid injection of sodium amytal for the lateralization of cerebral speech dominance. J. Neurosurg. 17, 266-282.

Wood, C. C., W. R. Goff, and R. S. Day. (1971) Auditory evoked potential during speech perception. Science 173, 1248-1251.

Phonetic Coding of Words in a Taxonomic Classification Task

G. Campbell Ellison*
Haskins Laboratories, New Haven, Conn.

That visually presented words are recoded into phonetic form is suggested by studies of two different types: those investigating short-term memory coding, and those using information processing techniques. Short-term memory studies (e.g., Conrad, 1964, 1972; Wickelgren, 1965, 1966; Baddeley, 1966) have shown that when nameable items must be remembered, those names are recoded into phonetic form. That is, when a list of phonetically similar items is presented visually for later recall, subjects perform more poorly than when the list is composed of phonetically dissimilar items.

It could reasonably be argued that visually presented words are recoded phonetically only when short-term memory is involved. Thus, to determine whether or not such recoding is applied to words because they are words, it is necessary to turn to procedures that avoid the use of short-term memory. Such is a usual property of information processing paradigms, of which we shall consider two pertinent examples.

It has been shown that when a subject is required to scan continuous text, crossing out *e*'s as he goes, he tends to miss those that are not pronounced (Corcoran, 1966). It is apparent from this that the whole word is processed before the target letter can be detected, that a phonetic code for the word is developed as a part of this processing, and that this development of the phonetic code interferes with performance.

Rubenstein, Lewis, and Rubenstein (1971) developed a task in which the subject is required to indicate whether each item he is shown is a word or not. They found that those nonwords that conformed to English spelling required more time to be classified than those that did not. Rubenstein et al. concluded that words and nonwords alike had to be recoded into phonetic form in order to access their representations in the lexicon (if indeed those representations exist).

There were some design and analysis flaws in this study, which were eliminated in an adaptation performed by Meyer, Schvaneveldt, and Ruddy (1974). They

---

*Also University of Connecticut, Storrs.

Acknowledgment: The author wishes to thank Michael Posner and Alvin Liberman for suggesting the basis for the paradigm used in this experiment, and the latter also for valuable comments on this research and on an earlier draft of this paper.

[HASKINS LABORATORIES: Status Report on Speech Research SR-41 (1975)]

presented pairs of items, both words and nonwords. They found that the time required to classify the items was affected by the graphic and phonetic regularity of the nonwords when these were present. Varying the graphic or phonetic similarity of a pair of items yielded appropriate variation in reaction time, viz. graphically similar items (words, primarily) that were phonetically distinct took longer to respond to, and vice versa, than fully similar items.

They concluded from this that a dual code (both phonetic and graphic in nature) is developed when a word or wordlike item is processed, and that such recoding is a necessary precursor to determining the meaning of a word. Like Rubenstein et al., they contend that words in the lexicon can only be accessed by the appropriately recoded representations, though they argue that such appropriate recoding is both graphic and phonetic in nature.

If this supposition is correct, it is clear that phonetic recoding of the stimulus is necessary to determine its meaning. However, it is possible that this recoding serves some other purpose, and that the lexicon is equipped to recognize words by their visual representations. The paradigm chosen for these studies may of itself require the use or development of a phonetic code, because the nonwords cannot be handled without using such a code. We cannot conclude that the subject is not able to determine something about the meaning of a word without first recoding it into phonetic form, or that an item cannot access its representation in the lexicon by its visual code. The extensive experience that adults have with visual representations of words argues for such an ability--to recognize words by the visual code.

We can apply the same argument to Corcoran's work. It may be that his procedure demands that the subject become aware of the identity of the word and the way it is constructed (phonetically) before any decision regarding its containing an e can be made. Awareness of the word's identity demands the production of a phonetic code.

If we are to make any decision about the use of a phonetic code in dealing with visually presented words, and, in particular, if we are to be able to apply such conclusions to reading, we must create procedures that make it highly unlikely that a phonetic code would be used in making whatever decision is required. As is the case with short-term memory coding research, there must be a penalty on the use of such a code. This is not sufficient, however, since that is what both Rubenstein et al. (1971) and Meyer et al. (1974) did in their procedures. We must, additionally, create a situation in which the subject can make extensive use of an alternate code--a code which, we may assume, is normally used in dealing with words. Since we are ultimately concerned with the possibility of generalizing the results of reading, it would also be proper to choose a paradigm that requires the subject to do something resembling what he does when he reads.

In the procedure to be used in this study, we ask, in effect, what else the subject knows about a word when he knows what it means. Specifically, he is asked to indicate whether or not each singly presented word he is shown belongs to a previously specified taxonomic category. If we use an easily defined category (such as the names of four-legged animals) and present as foil words homophones and rhymes of members of that category, as well as words visually similar to category members and words with little or no physical or phonetic resemblance, it should be possible to determine whether each word is recoded into phonetic

93

form prior to, or in the process of, determination of the meaning of that word. In such a case, we should expect to find that the subject takes longer to respond NO to words that are not animal names but that bear a strong phonetic resemblance to words that are.

Now, if it takes longer to respond NO to a foil word phonetically similar to a target than to one phonetically dissimilar to any in that category, we must conclude that a phonetic code is used in the performance of the task, and thus in determining the meaning of the word. If development of the phonetic code for a word does not occur in the process of ascertaining its meaning, there would be no increase in response latency; in this case, it could be concluded that phonetic recoding might occur, but that such a code does not enter into determining the meaning of the word. Naturally, any increase in reaction time to phonetically similar words implies only that such recoding of the word into phonetic form is involved in the determination of the word's meaning, and occurs before any decision is made regarding whether or not the word belongs to the category specified. It is not necessarily the case, as both Rubenstein et al. (1971) and Meyer et al. (1974) argue, that development of the phonetic code must precede finding the word in the lexicon.

To recapitulate, the subject's task is to make a keypress response to each word as it is presented visually, one response to targets, the other to foils. If the decision involves the use of the phonetic code of the item presented, reaction times to words phonetically similar to targets (i.e., rhymes and homophones) should be elevated with respect to those that are not. If only the visual appearance of a word is involved in such a decision (it clearly must play some role), then only foil words visually similar to targets should yield an increase in response latency. Overall, it is to the subject's advantage to make his decision based only on the visual appearance--the procedure allows him to do just that, if he is able.

METHOD

Apparatus

Subjects were run individually using a Lafayette-modified Kodak Carousel 750 slide projector--a projecting tachistoscope, and a reaction-time apparatus consisting of two keys connected to a relay, which in turn controlled a Lafayette Digital Stop Clock, so that depressing either key stopped the clock. The experimenter controlled presentation and clock onset by means of a key that controlled the relay and the tachistoscope via a Lafayette Decade Interval Timer.

Materials

Target stimuli were drawn from two categories: spelled-out numbers and four-footed animals (Battig and Montague, 1969). Four of each category were chosen to generate foils for the test group of stimuli, and two others of each to generate foils for the practice set. (See Appendix for complete lists.) For example, BEAR is a target. Its corresponding homophone is BARE, its rhyme is CARE, and the visually similar word is BEAT. Words in this last category were chosen to have the maximum number of letters in common with the corresponding target, with the identical letters as much as possible in the same position within the word. For example, BEAR and BEAT have three letters in common, and they are in the same position in both words. In all cases, the visually similar words have the same number of letters as their targets. The rhymes were chosen to have the least number of letters in common with the corresponding target, with both length and spelling being varied. Words of high frequency of occurrence (Thorndike and Lorge, 1943) were used as much as possible.

The result of this was a set of 72 words, which appeared twice each, for a total of 144. The first 48 constituted the practice set, and only reaction times for the remaining test words were used in analysis. The subject was not made aware of any distinction between the groups.

A full set of test words was composed of the following: 24 animal targets (12 words twice each), 24 number targets, and 8 each of homophones, rhymes, and visually similar words, for each target category. As each subject was told to target for only one category (and was not aware the other existed), there were thus 24 targets and 72 foils for each subject, of which 48 were theoretically neutral with respect to the target category.

In order to assess any effects due to a set for a particular visual pattern, as opposed to a graphic pattern, there were two complete sets of words, one wholly in uppercase letters, and the other with each word appearing once wholly in uppercase and once wholly in lowercase.

Subjects

Subjects were 30 University of Connecticut Introductory Psychology students, participating as part of a course requirement.

Procedure

The subject was seated at a desk on which the equipment rested, the two response keys in front of him. The experimenter sat where he could easily see which key was pressed on each trial, as incorrect responses had to be discarded before analysis. The subject was told that the purpose of the experiment was to see how quickly and accurately people could classify words as members or nonmembers of a given category. He was then told the category to target for.

The function of the apparatus was explained, and he was told to press the left-hand key for a nontarget word, and the right-hand key for a target. He was then given two practice trials, to become familiar with the operation of the equipment, prior to presentation of the full set. He was warned not to anticipate the classification of any word and that the results were of no use if inaccurate. The full set of 144 words was then presented, one at a time, with a break between the 72nd and 73rd words to allow the slide Carousel to be changed. Reaction time and hand used were recorded for each trial (word).

RESULTS

It was necessary first to discard protocols with error rates that were excessive. A 5 percent rate was established as an arbitrary criterion, and out of 96 test words, 8 errors was found to be the smallest whole number significantly greater than 5 by a χ^2 test ($\alpha=.05$). Correspondingly, a maximum of two errors was allowed on target items alone (this was necessary because of the disproportionate number of foils). That is, if a subject made more than two incorrect responses with the left hand, or more than seven incorrect responses overall, that subject's results were dropped from the analysis. Seven subjects were dropped for the former reason, and three for the latter.

This left 20 subjects, 10 of whom targeted for numbers, and 10 for animals. Of these, five each were given the uppercase set of words, and five each got the mixed set.

The data for the foils only were analyzed by a three-way analysis of variance, repeated measures on one factor (foil type--that is, homophones, rhymes, visually similar words, and neutral words). The reuslts are given in Table 1. There was no significant difference overall between the two target categories, though there is a trend in that direction (.25>p>.10). Nor was there any difference between the two sets of words, so font was not an effective variable. These are, in any case, of less interest and importance than the results in the lower portion of Table 1.

TABLE 1: Table of analysis of variance.

Source of variance	Sum of squares	Degrees freedom	Mean Square	F	
Target category (TC)	.0999	1	.0999	1.7343	NS
Font (F)	.0202	1	.0202	.3506	NS
Interaction (TC X F)	.0928	1	.0928	1.6111	NS
Between subjects	.9219	16	.0576		
Foil type (FT)	.1195	3	.0398	44.2222	.01
Interaction (TC X FT)	.0192	3	.0064	7.1111	.01
Interaction (F X FT)	.0068	3	.0023	2.5555	NS
Interaction (TC X F X FT)	.0046	3	.0015	1.6666	NS
Within subjects	.0466	48	.0009		
Total	1.3315	79			

That foil category is an effective variable indicates the usefulness of the procedure. Reaction time is clearly affected by the similarity of foils to their corresponding targets. What is more interesting, however, is the interaction between foil type and target category. The nature of this becomes clear when we consider Tables 2 and 3 together.

TABLE 2: Table of Newman-Keuls ordered differences among foil types for animal targets. (Scores shown are calculated values, not actual scores.)

Rhyme	Visual	Homophone	Foil type
3.9008 .01	9.2383 .01	13.0115 .01	Neutral
	5.3375 .01	9.1107 .01	Rhyme
		3.7732 .05	Visual

TABLE 3: Table of Newman-Keuls ordered differences among foil types for number targets. (Scores shown are calculated values, not actual scores.)

Rhyme	Homophone	Visual	Foil type
2.2943 NS	5.3069 .01	9.6466 .01	Neutral
	3.0216 .05	7.3523 .01	Rhyme
		4.3396 .01	Homophone

Clearly, the effect differs according to the target category. When a subject is required to target for the names of animals, reaction time is affected by both phonetic and visual similarity of the foil words to their corresponding targets. When the target category is spelled-out numbers, there is little or no effect due to phonetic similarity. The effect is, rather, a visual one.[1] Figure 1 shows the data in a more obvious way.

DISCUSSION

Not all of these results could have been anticipated, and they are consequently the more interesting. It is clear that two codes are involved in the

---

[1] This procedure has been replicated using the same categories and more extensive analysis, with identical results.

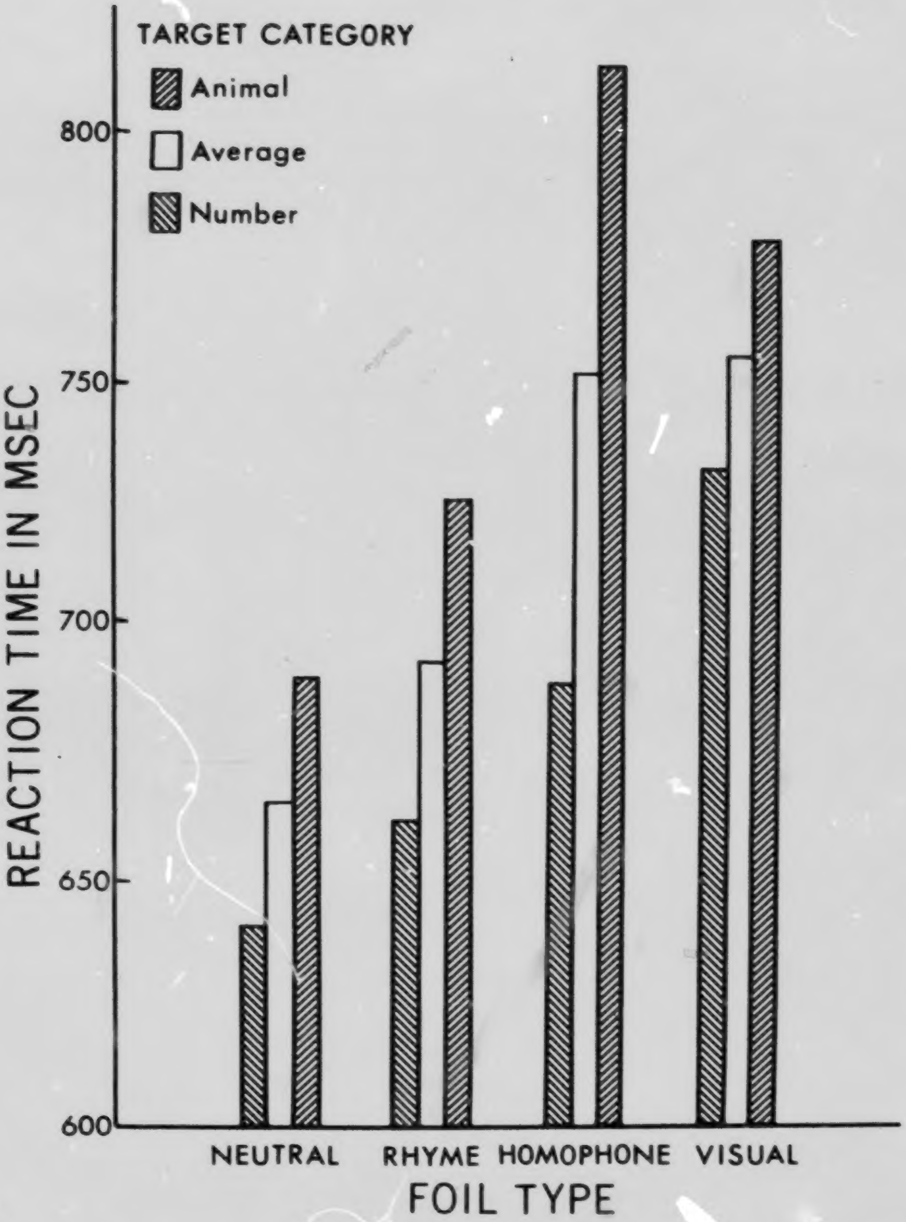

TARGET CATEGORY
▨ Animal
☐ Average
▧ Number
800
750
700
650
600
REACTION TIME IN MSEC
NEUTRAL
RHYME
HOMOPHONE
VISUAL
FOIL TYPE

FIGURE 1

task, both visual and phonetic. We can therefore conclude that a phonetic code is implicated in ascertaining the meanings of words.

The exception is that one does not deal with numbers in this way. The reasons for this may be crucial to a proper understanding of the results as a whole. First of these reasons is evidence that suggests a distinction between numbers and other verbal entities (e.g., nouns and adjectives). The former seem to be manipulated by right-hemisphere structures, as patients with certain kinds of temporo-parietal lesions in the nonlanguage hemisphere have great difficulty with numerical kinds of operations (e.g., Luria, 1966). There is a second aspect to this: the important operations one performs with numbers seem to have little or nothing to do with language.

A second reason, related to the first, is that numbers consist of a small set of logograms, and, even spelled-out, they may be conceived of as a set of logograms that is itself not very large. They are the only such set in English. It is reasonable that the printed or written versions of numbers be treated as logograms. In fact, it would be advantageous to learn to recognize the spelled-out version as the logographic version is recognized.

The third reason for assuming numbers to be different, even when spelled out, is that they form a closed and easily specified set. One knows immediately whether or not a word is the name of a number. The same cannot be said of a possible animal name.

For any or all of these reasons, we may suppose that it is very easy to respond merely to the visual form of a number. This might lead to shorter reaction times to words that are not the names of numbers, in the paradigm used in this study (which we have seen to be the case, though the difference was not significant). It is clear that a phonetic code must exist for the numbers. It also seems that the visual code is much more powerful, and overrides it.

Let us briefly consider these results in terms of the logogen theory of Morton (e.g., 1969). The logogen is a device (for want of a better word) that is specific to a particular item, such as a word, and responds to contextual, visual, and phonetic information regarding that item, making a response available when the sum of that information exceeds threshold. A logogen for a single word that is strongly specified in context might be very close to threshold, and so a response regarding it may be made much sooner than one that has to do with another word that is not so highly specified. Two factors that influence context are frequency of usage and the degree of specificity of the category to which the word belongs.

We must assume that development of the phonetic code takes time, and it may not begin until after the graphic code is developed (if a code is developed separately from the physical code in which the word is presented). A highly specified word, or one that is particularly used in a visual fashion, may exceed threshold before the phonetic code has developed far enough to affect the availability of the response.

If the response is not made available until after all the codes are fully developed, or if the specificity of the category does not lower the threshold of the logogen, we should expect to find that both phonetic and graphic similarities influence reaction time to foils.

We have then two explanations for the results. On the one hand, the names of numbers may be easier to specify, and the phonetic code may take too long to develop, thus reducing its effectiveness. On the other, the logogens appropriate to the names of numbers may react only or primarily to visual information, and not to the phonetic code.

It should be possible to examine this conflict between interpretations experimentally, by varying the frequency of occurrence of both targets and foils, and by creating small sets of targets that are memorized by the subject. If the former hypothesis is correct, that the threshold for number logogens is lower than for other words (such as animal names), then we should find that the condition in which subjects memorize the names of the targets and are given very high frequency targets and foils should yield results like those observed for the number foils in the present study. If, on the other hand, the latter hypothesis is correct, we should expect to find no difference, except for an overall reduction in reaction time to both targets and foils. Then we should conclude the results already observed to be due to differences in coding between numbers and other words.

REFERENCES

Baddeley, A. D. (1966) Short-term memory for word sequences as a function of acoustic, semantic, and formal similarity. Quart. J. Exp. Psychol. 18, 362-365.

Battig, W. F. and W. E. Montague. (1969) Category norms of verbal items in 56 categories: A replication and extension of the Connecticut category norms. J. Exp. Psychol. 80 (3, Pt. 2), 1-46.

Conrad, R. (1964) Acoustic confusions in immediate memory. Brit. J. Psychol. 55, 75-84.

Conrad, R. (1972) Speech and reading. In Language by Ear and by Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Corcoran, D. W. J. (1966) An acoustic factor in letter cancellation. Nature 210, 658.

Luria, A. R. (1966) Higher Cortical Functions in Man. (New York: Basic Books).

Meyer, D. E., R. W. Schvaneveldt, and M. G. Ruddy. (1974) Functions of graphemic and phonemic codes in visual word-recognition. Mem. Cog. 2, 309-321.

Morton, J. (1969) Interaction of information in word recognition. Psychol. Rev. 76, 165-178.

Rubenstein, H., S. S. Lewis, and M. A. Rubenstein. (1971) Evidence for phonemic recoding in visual word recognition. J. Verbal Learn. Verbal Behav. 10, 645-657.

Thorndike, E. L. and I. Lorge. (1943) The Teacher's Word Book of 30,000 Words. (New York: Teachers College Press).

Wickelgren, W. A. (1965) Distinctive features and errors in short-term memory for English vowels. J. Acoust. Soc. Amer. 38, 583-588.

Wickelgren, W. A. (1966) Distinctive features and errors in short-term memory for English consonants. J. Acoust. Soc. Amer. 39, 388-398.

APPENDIX

Animals

Targets	Homophones	Rhymes	Visuals
Horse	Hoarse	Course	House
Deer	Dear	Hear	Deep
Bear	Bare	Care	Beat
Hare	Hair	Fair	Harp

Numbers

Targets	Homophones	Rhymes	Visuals
One	Won	Done	Eon
Two	Too	Flew	Tow
Four	Fore	Bore	Foul
Eight	Ate	Late	Fight

Other Targets

Animals	Numbers
Sheep	Three
Mule	Seven
Cow	Thirty
Cat	Fifty
Lamb	Five
Wolf	Nine
Dog	Forty
Mouse	Sixty

On the Front Cavity Resonance, and Its Possible Role in Speech Perception

G. M. Kuhn
Haskins Laboratories, New Haven

ABSTRACT

Spectrographic data are presented which suggest that it may be possible to estimate the frequency of the fundamental resonance of the cavity behind the mouth opening, the "front cavity resonance," from information in the speech signal. It is shown that place of articulation information in the steady states, transitions, and bursts of F_2 (or sometimes F_3) can be reinterpreted to be information from the front cavity resonance. Furthermore, a number of synthesis results that have appeared anomalous when described in terms of numbered formants seem to find a coherent explanation in terms of the front cavity resonance. Implications for theories of speech perception include the possibility that an estimate of front cavity resonance frequency may serve for continuous articulatory reference.

INTRODUCTION

According to the acoustic theory of speech production, the fundamental resonance of the cavity next to the mouth opening, the "front cavity resonance," may be associated with any of the first four formants (Fant, 1960:72). But, as tongue constriction is relaxed, there is less dependence of any formant on one subpart of the vocal system, so little emphasis has been placed on cavity affiliations when describing the speech signal.[1] Instead, the description of acoustic cues for place of articulation remains largely in terms of numbered formants, with particular emphasis on F_2.

It is of interest, therefore, that the spectrographic data presented below suggest that it may be possible to estimate the front cavity resonance frequency from information in the speech signal. As a result, it appears that a more articulatory description of the acoustic cues can be provided, and that several anomalous results of experiments on acoustic cues can be explained.

---

[1] However, for a discussion of the effect of isolated articulatory movements on formant positions, see Delattre (1951).

Acknowledgment: F. S. Cooper, C. G. M. Fant, O. Fujimura, M. Studdert-Kennedy, A. M. Liberman, R. McGuire, P. Mermelstein, K. N. Stevens, and the Referee offered many helpful, substantive criticisms while this paper was in various stages of preparation. S. Koroluk and A. McKeon prepared the final manuscript and figures.

[HASKINS LABORATORIES: Status Report on Speech Research SR-41 (1975)]

The spectrographic data come from analysis of two types of speech. The first type is normal speech, and the second type is speech produced with a fricated source, or "fricative speech." In fricative speech, palatal frication is substituted for laryngeal voicing, and the nasal port is kept closed. The position of the palatal frication adjusts with the articulation until it feels more nearly velar in backed environments. It should be noted that the frication constriction is maintained even for speech sounds that are not normally characterized by significant constriction of the vocal tract (e.g., central vowels). Two interesting properties of fricative speech are, first, that it seems highly intelligible, and second, that the fundamental resonance of the front cavity appears as a prominent spectral component.[2] The acoustic similarities between fricative and normal speech suggest that a front cavity resonance frequency estimate can be made for normal speech.

On the Possibility of Estimating the Front Cavity Resonance Frequency

Figure 1 shows spectrographic analyses of the phrase "Where were you a year ago?," spoken under two conditions of excitation: fricative speech (top) and normal speech (bottom). Visual inspection of the top spectrogram indicates the presence of two components in the fricative speech token. The most obvious component varies in frequency from 700 to 3000 Hz and is visible in all excited portions of the token. Another component is fixed above 3500 Hz and is less visible when lip rounding increases. While the fixed component may be due to the fricative constriction, the variable component can be interpreted to be the fundamental, quarter-wave resonance of the front cavity. The variations in front cavity resonance frequency appear to reflect changes in the position of fricative constriction (from velar to prepalatal), and changes in lip opening (from rounded to retracted). Using the formula $l = c/4f$, and setting $c = 353$ m/sec (for 35°C), a quarter-wave resonance at 700 Hz would indicate that the front cavity has a functional length of about 12.6 cm; at 3000 Hz, a length of about 2.9 cm.

It comes as no surprise that the front cavity resonance should vary so continuously in fricative speech, since tongue constriction is extreme. What is interesting, however, is that this resonance can be traced so easily in the normal speech token. A comparison of the two spectrograms shows that this is the case. The comparison also illustrates the point that the fundamental resonance of the front cavity cannot always be associated with the same numbered formant: it may be associated with F_2 in /ɝ/ and /u/, but it is more strongly associated with F_3 in /i/.[3]

---

[2] We know of no reference to fricative speech in the acoustic phonetics literature. However, for a discussion relevant to fricative speech, see Fant (1960:72). There it is suggested that a static, three-section model of the vocal tract can show some of the essentials of velar and palatal articulation. Specifically, for the model of the articulation of /k/ or /g/ before /ɑ/, /æ/, or /i/, it is suggested that the fundamental resonance of the front cavity can be associated with F_2, F_3, or F_4, respectively.

[3] Sometimes the association of the front cavity with F_4 of /i/ is mentioned (Fant, 1960; see footnote 2, above), sometimes its association with F_3 of the same vowel (Fant and Pauli, 1974). What appears to have been the emphasis of the earlier discussion, and what we attempt to emphasize again here, is not the affiliation of the front cavity with a given formant, but the ability of the front cavity resonance to move more or less continuously in frequency given significant vocal-tract constriction.

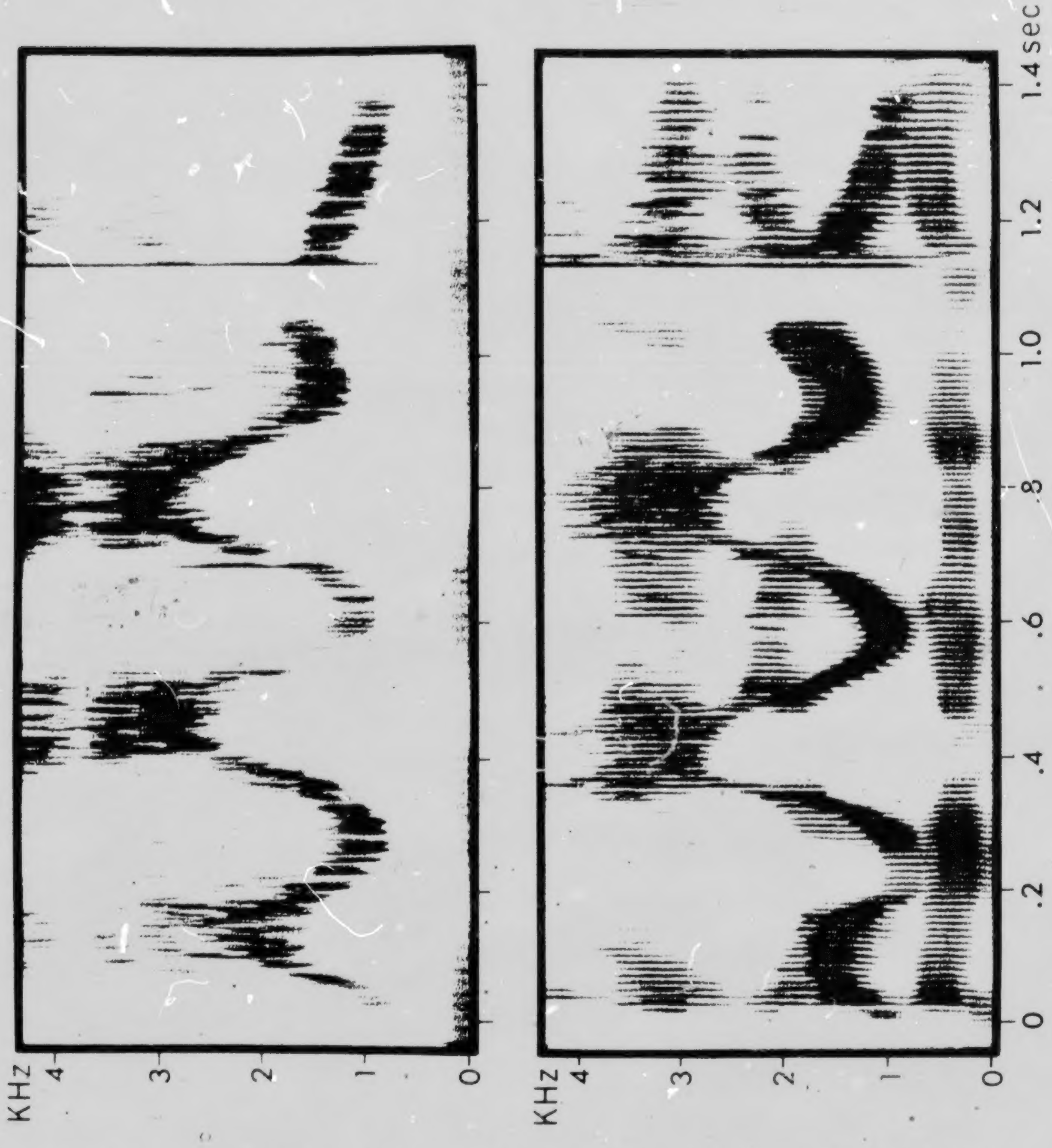


Figure 1: Spectrographic comparison of the phrase "Where were you a year ago?" for two conditions of excitation: fricative speech (left) and normal speech (right).

Figure 2 shows spectral cross sections of eight vowels, all spoken by the same adult male. There are two sections per vowel, one each from fricative speech (left) and normal speech (right).

It may not be inappropriate, at this point, to insert a comment about the ease of production of these fricative speech vowels. Fant (1960:115) reports vocal-tract cross-sectional areas for /i e ɑ o u/. In the region of the tongue constriction, the cross-sectional area appears to fall to 1 cm^2 or less for /i ɑ o u/, but to no less than 2 cm^2 for /e/. Similarly, it seems easy to make the constriction for a satisfactory fricative speech close front vowel (here, /i/ and /I/). It also seems easy to make the constriction for the vowels with a backed tongue position (/a ʌ U u/), where we were more aware of manipulating the lip opening when trying to adjust the perceived color. However, it seems less easy to lower the jaw and produce convincing fronted palatal constriction for the more open front vowels /ɛ/ and /æ/.

These cross sections give further indication that a front cavity resonance frequency estimate can be made for normal speech. In these sections, the length of the front cavity seems to have an important effect on the overall spectral shape. The fricative and normal speech spectra seem to be shaped toward the high frequencies when the front cavity is short, as for /i/, and toward progressively lower frequencies as the front cavity is apparently lengthened for each successive vowel. In addition to the effect of the length of the front cavity, there also seems to be an effect due to the amount of tongue constriction involved. The greater the constriction, the more the front cavity resonance in the fricative speech seems to correspond to a formant in the normal speech. This correspondence seems very close for F_3 of /i/ and /I/, and for F_2 of /a ʌ U u/. For all eight vowels, however, the front cavity seems to be associated with what is perhaps the most intense group of formants: with the F_3 group for /i I ɛ æ /, and with the F_2 group for /a ʌ U u/.[4] Notice in particular the change in overall spectral shape from /æ / to /a/, where the front cavity shifts its strong association from F_3 to F_2 and the weight of the spectrum shifts to frequencies below 2000 Hz.[5] This change occurs despite the fact that the frequencies of F_1, F_3, and F_4 are essentially unchanged. These comparisons with

---

[4] Such phrases as "most obvious component" or "perhaps the most intense group of formants" should be accepted only with qualification. Figures 1, 2, and 3 show speech spectra after lift has been applied (approximately 6 dB per octave between 300 and 3000 Hz). Also, in Figures 1 and 3, automatic gain control and 300 Hz "broadband" filtering have been applied. These operations have been made available on commercial sound spectrographs because they have been thought helpful for revealing perceptually relevant aspects of speech. This is not enough, of course, to make us want to assume that such operations make speech spectrograms look exactly like speech sounds.

[5] The front cavity resonance in fricative speech appears to be most closely associated with F_3 of /i I ɛ æ / and with F_2 of /a ʌ ɑ U ɔ u/. This association is consistent with the nomograms of Figure 1.4-9 of Fant (1960), where the cavity affiliations of F_2 and F_3 appear to change at about 2000 Hz. For the model, this change has a constriction coordinate of approximately 11 cm from the glottis, which, in turn, is consistent with the estimate of "two-thirds of the total length of the vocal-tract" of Stevens and House (1956).

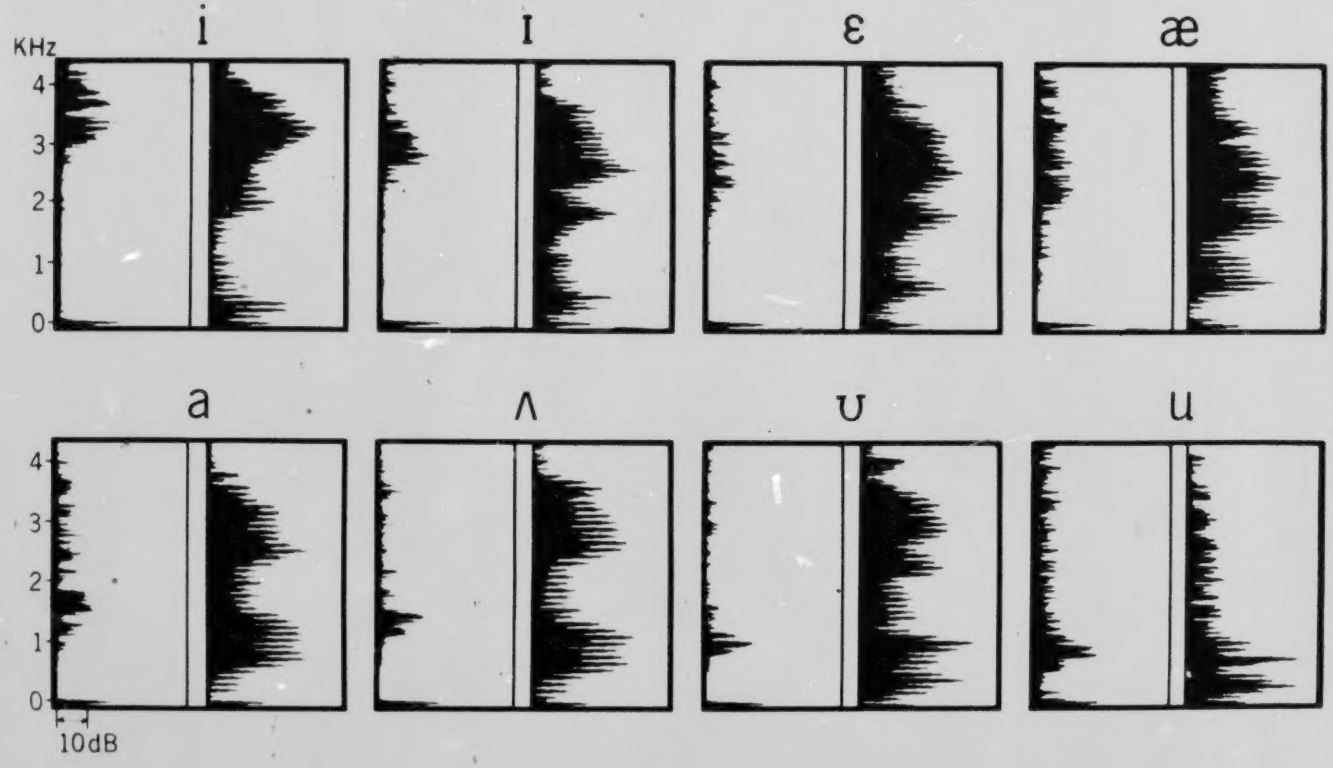


Figure 2: Spectral cross sections of eight vowels. There are two sections per vowel, one each from fricative speech (left) and normal speech (right).

fricative speech seem to lead us to an observation about spectral shape that is substantially the same as that made by Fant (1960:123), namely, that the front cavity can have an important effect on F_2 (and thus on the mean of F_1 and F_2), or on the mean of F_2 and all higher formants.

Figure 3 shows spectrograms of 12 consonant-vowel syllables, the consonants /b d g/ followed by the vowels /i æ ɑ u/. There are two spectrograms per syllable, one each from fricative speech (left) and normal speech (right). These spectrograms indicate that a front cavity resonance frequency estimate can be made for highly constricted normal speech consonants. They show the remarkable similarity of burst and transition information in fricative and normal speech. Notice again the shift in spectral weight toward the lower frequencies, this time as the vowel goes from /æ/ to /ɑ/.

These observations suggest a general effect of the front cavity, that it is a determiner of the overall spectral shape. Nevertheless, it appears possible to construct a formula to estimate the front cavity resonance frequency from formant frequency data. For constricted vowels, this formula should place the front cavity resonance frequency estimate somewhere between the low values for F_2, as in back vowels, and the high values of F_3, as in front unrounded vowels like /i/.

Carlson, Fant, and Granström (1973) have expressed exactly these concerns in designing a formula for predicting a perceptual "F_2 prime" for vowels. The notion of F_2' arises from a desire to represent natural vowels in a perceptually equivalent two-formant space (see, e.g., Delattre, Liberman, and Cooper, 1951; Fant, 1959). The F_1 of the natural vowel is replaced by the F_1 of the two-formant equivalent, while all higher formants of the natural vowel are replaced by the F_2 (the so-called F_2') of the two-formant equivalent. From the data of a matching experiment in which techniques for two-formant, parallel resonance synthesis were used, Carlson, Granström, and Fant (1970) report values of F_2' for several Swedish vowels. The matching experiment values of F_2' range from about 700 Hz for /u/ to about 3000 Hz for /i/. These limiting values, and the other, intermediate values reported, appear to lie close to the front cavity resonance frequency as estimated from fricative speech. The thought arises, then, that the front cavity resonance frequency may be what F_2' predicts. If this is so, then it might be appropriate to estimate the front cavity resonance frequency using the formula proposed by Carlson et al. (1973). That formula is

$$F_2' = \frac{F_2 + c(F_3F_4)^{1/2}}{1 + c}$$

where

$$c = \left(\frac{F_1}{500}\right)^2 \left(\frac{F_2 - F_1}{F_4 - F_3}\right)^4 \left(\frac{F_3 - F_2}{F_3 - F_1}\right)^2$$

The formula apparently generates the results of the matching experiment to within 65 Hz, on the average. Carlson et al. (1973) report that the values of F_2' predicted by the formula are also within 75 Hz, on the average, of values predicted by a model of the cochlea. When the reference vowels from the matching experiment were the input to their cochlear model, then the two most prominent peaks in the output "were found to correspond closely to F_1 and the F_2' of the

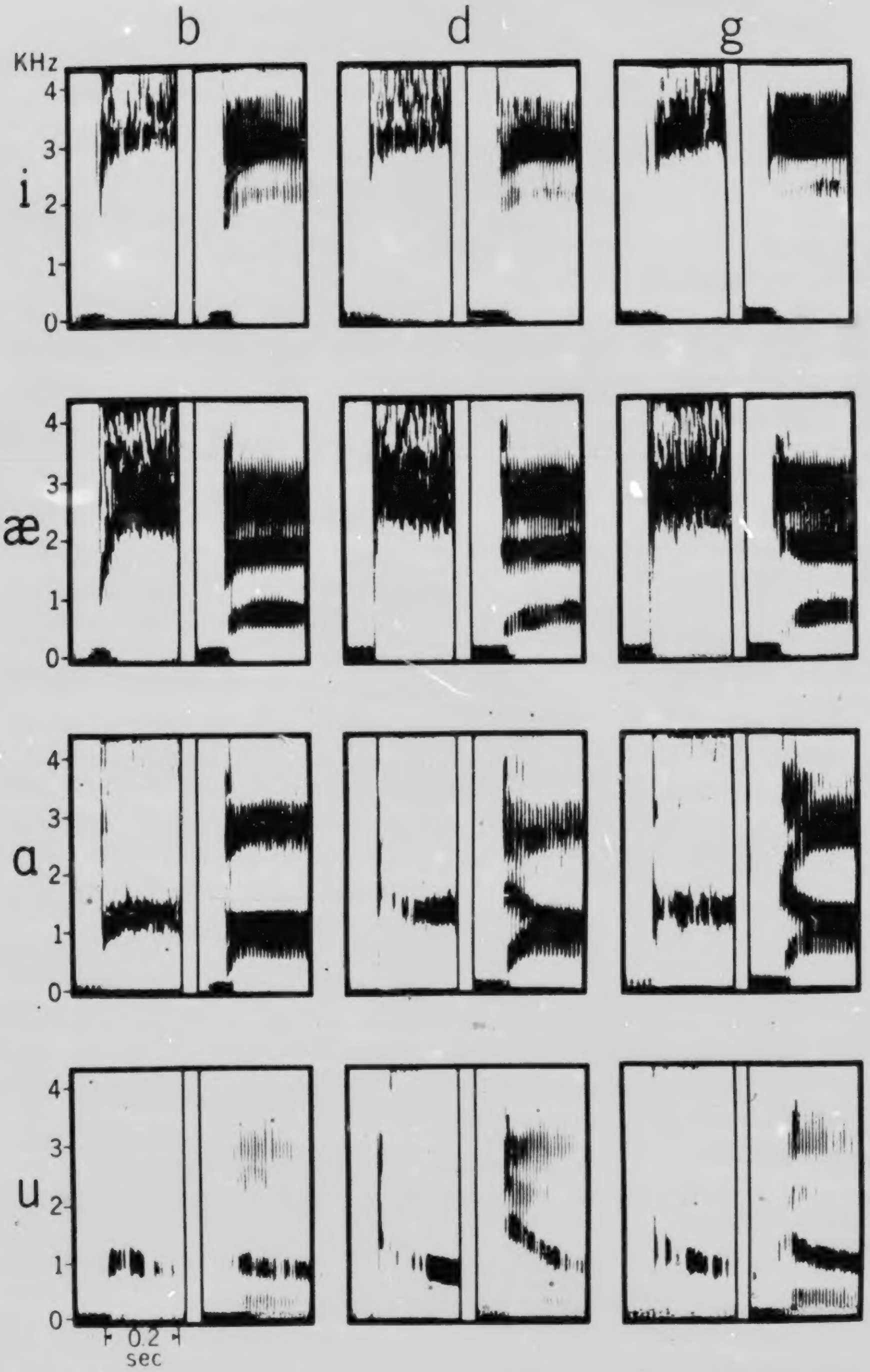


Figure 3 : Spectrograms of 12 CV syllables, the consonants /b d g/ followed by the vowels /i æ ɑ u/. There are two spectrograms per syllable, one each from fricative speech (left) and normal speech (right).

two-formant matching." Thus, there is some rather indirect evidence that the front cavity resonance frequency could also be estimated from speech data that is in a form perhaps more like that found in the auditory system. The authors attribute the close agreement between the formant equation and those of the cochlear model, at least in part, to "single component prominence." This explanation appears to be consistent with an emphasis on the front cavity as a determiner of the overall spectral shape.

The methods for predicting F_2' suggest how the front cavity resonance might possibly be estimated for vocalic sounds. This estimate might be expected to be more consistent for the more constricted sounds, where the formant frequencies can move more continuously. In return, the front cavity resonance may provide an articulatory rationale for F_2', which, heretofore, has been motivated mainly by perceptual considerations.

A Reinterpretation of Cues from F_2 and F_3

Since it appears that the front cavity resonance could be estimated from information that is intense in the speech signal, one may ask for indications that this happens, in fact, during speech perception. What follows is an attempt to reinterpret familiar data from studies of the perception of synthetic speech, in a fashion consistent with a possible role for the front cavity resonance.

The front cavity resonance appears to play a role in vowel perception. Single-formant equivalents of two-formant vowels have been reported for /i u o ɔ ɑ a/ (Delattre, Liberman, Cooper, and Gerstman, 1952). These single-formant equivalents lie close to the frequency of the front cavity resonance as estimated from fricative speech. In two-formant vowel synthesis, weighted averaging of the natural F_2 and F_3 has been used for front vowels, where the front cavity resonance in the natural case may be more strongly associated with F_3. For example, the two-formant /i/ of Delattre et al. (1952) had an F_2 at 2880 Hz, and that of Liberman, Delattre, Cooper, and Gerstman (1954) had an F_2 at 2760 Hz, whereas the natural F_2 appears to be located at about 2300 Hz and the natural F_3 at 3000 Hz (Peterson and Barney, 1952). Again, Carlson et al. (1970, 1973) are investigating a perceptual F_2' that, for both front and back vowels, may track the front cavity resonance.

The front cavity resonance appears to play a role in the perception of stop consonant formant transitions. The front cavity resonance in fricative speech is close to F_2 in /a/, and Liberman et al. (1954) found that changes in the F_2 transitions alone were sufficient to produce /ba/, /da/, and /ga/ responses. But the front cavity resonance is close to F_3 in /i/, and Harris, Hoffman, Liberman, Delattre, and Cooper (1958) produced /bi/, /di/, and /gi/ responses by changing the F_3 transitions alone.[6]

Finally, the front cavity resonance appears to play a role in the perception of stop consonant bursts. Only the voiceless stop bursts are mentioned here, because of the relevance of the cited synthesis results. The descriptions appear to be applicable to the homorganic voiced stops.

---

[6] This is our interpretation of their results. Harris et al. (1958) showed that a flat F_2 and different rising transitions of F_3 could cue /gi/ and /di/ responses. They also showed that a sharp rise in both F_2 and F_3 could cue a /bi/ response. We are interpreting this last case to be equivalent to an F_3 transition that starts below a flat F_2.

For /t/ bursts, the cavity behind the mouth opening is small, extending back only to the alveolar constriction, regardless of the resonator configuration for a following vowel. Synthesis should therefore reveal the importance of a high frequency, and relatively unchanging spectral component. It does: Liberman, Delattre, and Cooper (1952) obtained /t/ responses for bursts above 3000 Hz before the vowels /i e ɛ a ɔ o u/.

For /p/ bursts, the spectrum immediately following lip release can be broad and flat, because there is no resonator of significance in front of the constriction. Then, as the lips open further, the front cavity resonance can rise abruptly in frequency and amplitude. Before unrounded vowels, these excursions may be quite salient, but before rounded vowels, if the lip opening remains small, they would be diminished. (Compare the spectrograms for /bi/ and /bu/ in Figure 3.) In synthesis, the excursions of the front cavity resonance might therefore be expected to play a more important role before unrounded vowels than before rounded ones. Indeed, Liberman et al. (1952) found that /p/ responses dominated when a schematic burst was positioned some 360 Hz below the formant closest in frequency to the front cavity resonance in a following /i e ɛ a/. But before /ɔ o u/, they found /p/ responses to dominate when the burst was neither near the frequency of the front cavity resonance, nor in the /t/ region, but rather around 1500 Hz.

For /k/ bursts, the cavity behind the mouth opening extends back to the hump of the tongue, so that a front cavity resonance component of the burst should be affected at once by concomitant positioning of the tongue hump for a following vowel. The question is whether synthesis reveals a strong dependence of the burst on the frequency of the formant closest in frequency to the front cavity resonance of a following vowel. In fact, the data of Liberman et al. (1952) show that /k/ responses predominated when a schematic burst was placed at, or slightly above, the formant closest in frequency to the front cavity resonance in a following /i e ɛ a ɔ o u/.

An Explanation of Anomalies

Stevens and House (1956) suggested that some anomalies encountered in perceptual studies of transitional cues may be attributed to the changing cavity affiliations of F_2 and F_3. Since the possibility of explaining anomalies is at least as compelling as that of reinterpreting phenomena, it is interesting to note that if the role of the front cavity resonance is emphasized in describing the speech signal, several anomalies of the acoustic phonetic literature seem to find an explanation. Consider the explanations of the following three anomalies in terms of a possible role for the front cavity resonance.

One anomaly is the burst of noise at 1440 Hz that cued a /pi/, /ka/, or /pu/ response in Liberman et al. (1952). Before /i/ this burst appears to be interpreted as part of the rise in frequency of the front cavity resonance as it moves up to F_3. Before /a/, the burst appears to be interpreted as part of the fall in frequency of the front cavity resonance as it moves to a slightly lower value in F_2. Before /u/, it appears to be interpreted as part of a flat, lip-release spectrum and was a somewhat weaker cue. The /pi/-/ka/-/pu/ result is then consistent with the suggestion that the front cavity resonance plays a role in the perception of /p/ and /k/.

A second anomaly is the F_3 transition that was important for the /d/ in /di/ but not for the /d/ in /du/ (Harris et al., 1958). This result may be due to the fact that the fundamental resonance of the front cavity is strongly associated with F_3 of /i/, but not with F_3 of /u/. If so, the /di/-/du/ result suggests that transitions of the front cavity resonance can play a role in the perception of /d/.

A third anomaly is the F_2 transitions in two-formant synthesis of /g/: one could extrapolate the F_2 transitions to a virtual /g/ locus at 3000 Hz before /i e ɛ a/, but before /ɔ o u/ the locus would have to be much lower in frequency, if indeed it existed at all (Liberman, 1957). Figure 3, above, indicates that, like the velar bursts, velar transitions covary in frequency with the front cavity resonance of the following vowel. Indeed, the transitions that produced predominantly /g/ responses in Liberman et al. (1954) lie at the same frequency as bursts that produced predominantly /k/ responses in Liberman et al. (1952). These results suggest that the real, relative frequency of the transition of the front cavity resonance plays a role in the unchanging perception of the velar stop consonants.

DISCUSSION

Acoustic anomalies like those above led Liberman, Shankweiler, and Studdert-Kennedy (1967) to express the belief that speech perception might involve a simplifying reference to articulation. The data and arguments of this paper suggest that such a simplifying reference may be available directly from the speech signal: despite the acoustic complexity of the anomalies mentioned, an interpretation in terms of the front cavity resonance seems to provide, in each case, a rational account.

One can try to show that an articulatory reference is available in the speech signal without arguing that this reference is interpreted by a process of analysis-by-synthesis. The task of synthesizing an acoustic pattern to subtract from the incoming signal now appears simpler: the rules required to generate those curious speech acoustics do not seem anomalous when expressed in terms of the resonator system that produces them. But at the same time, the task of directly perceiving the incoming signal appears simpler, too: there appears to be an intense component of the signal that carries important information about place of articulation.

These observations suggest that a person who is perceiving speech might be described as one who is interpreting at least part of the signal as a contribution specifically of the front cavity. Given the quarter-wave resonator model, a front cavity resonance frequency estimate is also an estimate of the front cavity length. And for a given articulation, the front cavity length may not vary a great deal across individuals, not, for example, as much as the length of the pharyngeal cavity (Fant, 1966). Therefore, a front cavity resonance frequency estimate would be almost an estimate of place of articulation. It is necessary to say "almost" an estimate of place of articulation for at least two reasons: first, because of possible differences in front cavity length; and second, because similar lengths of the front cavity could arise in different combinations of fronted tongue constriction with lip rounding, or backed constriction without rounding. This last consideration indicates a possibly important use for continuous tracking of the front cavity resonance: spectra that are articulatorily ambiguous might be disambiguated if the preceding or following configuration of the slowly changing resonator system is unambiguous.

CONCLUSION

We have attempted to present a new technique (fricative speech) and articulatory rationalizations of some of the acoustic cues for speech. These have been used to emphasize a relationship that seems to deserve more attention, namely, the relationship between the fundamental resonance of the front cavity and the perceived place of articulation. This relationship would tend to arise to the extent that speech requires significant constriction of the vocal tract, as may be the case for consonants generally, and for many (though not all) vowels. We believe that such constriction contributes to the solution of the problem of deriving an articulatory description from the acoustics of speech. A front cavity resonance frequency estimate seems to be a useful way to represent part of that contribution.

REFERENCES

Carlson, R., G. Fant, and B. Granström. (1973) Two-formant models, pitch, and vowel perception. Paper presented at the Symposium on Auditory Analysis and Perception of Speech, 21-24 August, Leningrad.

Carlson, R., B. Granström, and G. Fant. (1970) Some studies concerning perception of isolated vowels. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) QPSR-2/3, 19-35.

Delattre, P. (1951) The physiological interpretation of sound spectrograms. PMLA 66, 864-875.

Delattre, P., A. M. Liberman, and F. S. Cooper. (1951) Two formant synthetic vowels and cardinal vowels. Le Maître Phonétique, July-December.

Delattre, P., A. M. Liberman, F. S. Cooper, and L. J. Gerstman. (1952) An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesized from spectrographic patterns. Word 8, 195-210.

Fant, C. G. M. (1959) Acoustic analysis and synthesis of speech with applications to Swedish. Ericsson Tech. 1, 3-108.

Fant, C. G. M. (1960) Acoustic Theory of Speech Production, 2nd ed., 1970. (The Hague: Mouton).

Fant, C. G. M. (1966) A note on vocal tract size factors and nonuniform F-pattern scalings. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) QPSR-4.

Fant, C. G. M. and S. Pauli. (1974) Spatial characteristics of vocal-tract resonance modes. Preprints of the Speech Communication Seminar, Stockholm. 1, 121-132.

Harris, K. S., H. Hoffman, A. M. Liberman, P. Delattre, and F. S. Cooper. (1958) Effect of third-formant transitions on the perception of the voiced stop consonants. J. Acoust. Soc. Amer. 30, 122-126.

Liberman, A. M. (1957) Some results of research on speech perception. J. Acoust. Soc. Amer. 29, 117-123.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Liberman, A. M., P. Delattre, and F. S. Cooper. (1952) The role of selected stimulus-variables in the perception of the unvoiced stop consonants. Amer. J. Psychol. 65, 497-516.

Liberman, A. M., P. Delattre, F. S. Cooper, and L. J. Gerstman. (1954) The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monogr. 68, 1-13.

Peterson, G. E. and H. L. Barney. (1952) Control methods used in a study of the vowels. J. Acoust. Soc. Amer. 24, 175-184.

Stevens, K. N. and A. S. House. (1956) Studies of formant transitions using a vocal-tract analog. J. Acoust. Soc. Amer. 28, 578-585.

Synthetic Speech Comprehension: A Comparison of Listener Performances with and Preferences Among Different Speech Forms

Patrick W. Nye, Frances Ingemann,* and Lea Donald+
Haskins Laboratories, New Haven, Conn.

ABSTRACT

Two passages of text obtained from a reading test were converted into phonetic strings; initially by machine and later by hand. Subsequently, these phonetic texts were input to a selection of synthesis-by-rule algorithms that have been developed at Haskins Laboratories during the past four years. Two groups of subjects heard one of the text passages in natural speech and the other text in one of four alternate forms of synthetic speech. After each hearing, the subjects were timed, under self-paced conditions, as they answered questionnaires designed to assess their comprehension. The results show that the subjects' comprehension expressed in terms of the time taken to complete the questionnaire improved with successive synthesis algorithms. In addition, the hand-prepared texts contributed to better performances and the natural speech proved superior but by a relatively small amount.

In a second experiment, samples of all four synthetic speech forms were presented in pairs and the same subjects were asked to identify the speech form they preferred. An examination of these data show that the subjects' preferences ranked in the same order as did their performances in the previous experiment.

INTRODUCTION

A listener's ability to comprehend the contents of a passage read aloud depends on a number of factors. Many of these are closely interrelated, although, for the purposes of this discussion, they will be considered separately. For example, intelligibility is a factor that is frequently assessed by examining the responses of listeners to words or syllables delivered in isolation. However, although the results have an obvious bearing on comprehension, the extrapolation of these data to predict general comprehensibility is an uncertain art because of the difficulties of accounting for prose style and content. A second factor in comprehension concerns the prosodic patterns of the speaker's delivery--the speaker's usage of loudness, voice pitch, duration, and overall

---

*Also Department of Linguistics, University of Kansas, Lawrence.

+Also Department of Linguistics, University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-41 (1975)]

speaking rate. Last, there are such factors as the speaker's accent or dialect, and the characteristics of the amplifying or reproducing equipment if any is involved.

Listeners such as the blind, many of whom depend almost entirely on speech as a means of acquiring information, are particularly concerned about speaker "quality"--where the term quality is used in a broad sense to cover all the prosodic factors cited above. However, the availability of good quality readers, particularly volunteers, is restricted, and for this and many other reasons the process of producing spoken recordings for the blind is extremely slow. Several months can elapse between the publication of a new book or periodical and its availability in spoken form to blind subscribers. It is this fact that argues most strongly the need for an automatic reading system.

As a part of basic research on speech, Haskins Laboratories have been working for several years on the development of a Reading Machine for use by the blind and reading handicapped. During this period, with the objective of improving some aspect of speech quality, several different versions of the Synthesis-by-Rule program (Mattingly, 1968) have been designed by Kuhn to control two types of synthesizer--one of the Laboratories' own design and the other an OVE-III (Liljencrants, 1968). Using these programs, a prototype Reading Machine system has been assembled (Cooper, Gaitenby, Mattingly, Nye, and Sholes, 1972) that is capable of reading typewritten texts and converting them to synthetic speech with only occasional editorial intervention by a human operator. During the past two years, the Laboratories have been conducting evaluation studies to assess the quality of synthetic speech and to determine its potential for early application to the problem of providing blind people with faster access to printed information. Work reported in previous Status Reports has been concerned with measurements of the intelligibility of synthetic speech (Nye and Gaitenby, 1973, 1974) and the results of these studies have pointed out that several synthetic phonemes--particularly the fricatives--are poorly identified compared with their counterparts in natural speech. From these data, which yield error rates differing by as much as a factor of 10, it is apparent that, a priori, one could expect that a listener's comprehension of synthetic speech would lie below his comprehension of natural speech. However, there still remain the crucial questions: "Compared with natural speech, how good is the comprehension of a long text where natural redundancy is likely to compensate for losses in phonetic intelligibility?" and "Will blind listeners be tolerant of the deficiencies of synthetic speech in return for faster access to printed matter?"

To begin answering the first of these questions, a simple experiment based on a reading test was designed to derive a measure of the comprehensibility of synthetic versus naturally spoken text passages. Secondary objectives were (1) to determine the degree of improvement in speech quality contributed by successive speech synthesis programs and different synthesizers, (2) to assess the relative performance of synthesis programs using hand-edited versus purely automatically derived phonetic input, and (3) to compare the comprehensibility measurements with the results of a speech quality preference test.

METHOD

The reading test was designed to compare the comprehensibility of texts generated by three synthesis programs, employing two different synthesizers and two sources of phonetic input. The synthesis programs differed from one another in terms either of the tabular phonetic values used or the calculations that they performed to derive the control parameters fed to one of the two speech synthesizers.

Two passages of text were selected from a published reading test (Raygor, 1970) intended for college-bound and college students. The texts were matched for reading difficulty and were both on the subject of "tunnels." Text A contained roughly 2000 words, while text B contained about 1700 words. Copies of both texts were then converted from their orthographic form into phonetic strings by means of the Reading Machine program. No human intervention beyond ensuring that all the necessary words were contained in the computer-stored dictionary was involved. Phonetic transcriptions of the same two texts were also prepared by a linguist to represent, within the limitations of the OVEBORD or JUN74 program (see Ingemann, 1975), the way each sentence might be spoken. The two input strings differed principally in the placement of prosodic markers and the use of reduced versus full forms. Finally, these input strings were presented to the three synthesis programs and their associated synthesizers, which differed primarily in the circuitry of their formant resonators: in the first, an OVE-III, the resonators are connected in series; whereas in the older Haskins Laboratories synthesizer the resonators are connected in parallel. To limit the scale of the experiment to a manageable size, only a selected number of the possible "synthesis combinations" (i.e., combinations of algorithms, synthesizer, and text) were examined. These different speech forms are identified as follows:

1. DEC71-HO = algorithm: Slightly modified version of Mattingly (1968)
   synthesizer: Haskins Laboratories parallel formant synthesizer
   rules: Kuhn, available in December 1971
   text: Automatically derived phonetics

2. DEC73-OO = algorithm: Designed by Kuhn in 1973 (see Kuhn, 1973)
   synthesizer: OVE-III serial synthesizer
   rules: Kuhn, available in December 1973
   text: Automatically derived phonetics

3. DEC73-OE = algorithm: As above, designed by Kuhn in 1973
   synthesizer: OVE-III serial synthesizer
   rules: Kuhn, available in December 1973
   text: Hand-edited phonetics prepared by Ingemann

4. JUN74-OE = algorithm: Slightly modified version of Kuhn (1973)
   synthesizer: OVE-III serial synthesizer
   rules: Ingemann, available in June 1974 (see Ingemann, 1975)
   text: Hand-edited phonetics prepared by Ingemann

Each of these "synthesis combinations" receiving input from both texts A and B yielded a total of eight recordings. The speaking rate varied slightly among the different synthesis routines from a low of 133 words per minute (wpm) (DEC71-HO) to a maximum of 154 wpm (JUN74-OI). To provide a "control" condition, natural speech recordings of texts A and B were made by a male speaker in a moderate New York dialect that was fully familiar to all the listeners who completed the test. The speaking rate was 170 wpm.

Twenty-four college students were employed as "experimental" listeners for a fixed sum. Half of the students heard text A in one of the four forms of synthetic speech and then text B in natural speech. The remainder heard text B in synthetic speech and text A spoken naturally. Text A, in either synthetic or normal speech form, was always heard before text B. (A natural speech pilot experiment in which text B preceded text A for half of the trials provided no evidence that the order of presentation had any bearing on the difficulty that a subject experienced on a particular text.) The combinations of text and synthetic speech form that were assigned to individual listeners are shown in Table 1.

TABLE 1

Subject Numbers	Text	Speech Form
1 - 3	A	DEC71 - HO
4 - 6	B	DEC71 - HO
7 - 9	A	DEC73 - OO
10 - 12	B	DEC73 - OO
13 - 15	A	DEC73 - OE
16 - 18	B	DEC73 - OE
19 - 21	A	JUN74 - OE
22 - 24	B	JUN74 - OE

After hearing a text played through once without interruption, the listeners were required to answer 14 multiple-choice questions. These questions sought factual information from the texts and offered four possible answers to each question. One question on each text was concerned with numerical data, and a further 10 questions required answers that were either direct quotations or close paraphrases of short statements contained in the text. Answers to the remaining four questions were less direct and required the synthesis of facts distributed over a paragraph of text (average length of about 50 words).

Two factors were assumed to govern the listeners' performances on the questionnaire: the degree to which they had succeeded in interpreting and understanding the speech content and the amount of prior knowledge they may have had about the subject matter. With the objective of assessing the prior knowledge factor, the two questionnaires were presented to a new group of 12 student "readers" who, without hearing the texts, attempted to select the most plausible answer to each question or, failing that, picked an answer at random. These students were of academic status and background comparable to those of the "experimental" listeners.

The results of the prior knowledge test are shown in Table 2. Adopting the null hypothesis that all of the answers were selected at random, the binomial distribution was used to predict the number of students who could be expected to select correctly the answers of up to 8 questions out of the total of 14. These predicted data also appear in Table 2. To test the hypothesis, a χ^2 test was made of the expected numbers versus the actual numbers of students choosing correct answers. The result indicated that the actual data are consistent with the null hypothesis at a confidence level in excess of 5 percent. Thus, the phrasing of the questions or the reader's general knowledge provided very little help in choosing the correct answers.

TABLE 2

Number of correct answers chosen by a student (No)	Number of students who chose No answers		Frequency of students choosing No answers (predicted by binomial distribution)
	Text A	Text B	
0	0	1	0.3
1	0	2	1.2
2	3	1	2.5
3	1	6	3.4
4	5	1	3.1
5	0	0	2.1
6	3	1	1.0
7	0	0	0.4

Text A, $\chi^2 = 10.9$
Text B, $\chi^2 = 8.5$ (6 degrees of freedom)

Each student from the "experimental" group listened to the recordings in the presence of an experimenter equipped with stopwatch. At the end of each recording the stopwatch was started and the listeners immediately turned their attention to the questions and answered them at their own pace. However, in nearly all instances, at the end of one pass, some of the questions were left unanswered. After noting the time that had elapsed up to that point (T_1) and after rewinding the tape, the stopwatch was restarted. The listeners were then allowed selectively to replay passages and check off answers until they were confident that all the questions had been answered correctly. The time taken in this second phase of question-answering was also recorded (T_2). Exactly the same procedure was followed for text B.

Upon completing the answers for both texts, each listener was given a short passage in two synthetic speech forms and asked to state which he or she preferred. All possible pairings of the four speech forms were examined and their relative distance on an arbitrary preference scale (labeled from 0 to 7) was computed by the method of pair comparisons (Guilford, 1954).

RESULTS

The goals of the data analysis were to assess listeners' performances on synthetic and naturally spoken texts and their preferences among different speech forms. Tests for these differences were made statistically. Once again, in accordance with basic principles, a null hypothesis was adopted, namely, that the data were drawn from the same distribution, i.e., no differences were anticipated. However, individual differences in listening skills were likely, and their effect was offset where possible by applying tests to differences between individual performances with synthetic and natural speech.

Differences Between Synthetic and Natural Speech

An analysis of the observations listed in Table 3 reveals that regarding the time T_1 taken to complete the first pass through the questionnaires, the null hypothesis is confirmed and no differences emerge between pooled synthetic and natural data. However, the same treatment applied to T_2 shows that the second period (needed to complete the questionnaire) is an average of 4.5 minutes in length for natural speech and 1 minute and 45 seconds longer when the listener works with a synthetic speech text. The probability that this difference arises by chance is small ($p \approx 0.025$) and suggests that the listener requires 23 percent more time to understand the synthetic speech passage. A comparison of the number of erroneous answers in the two conditions shows, however, no significant differences. This finding was not unexpected because the instructions given to the listeners stressed that they were to continue working until they were satisfied that all of their answers were correct. Thus, verification of the null hypothesis in this case merely indicates that the listeners followed their instructions with equal consistency in the two conditions.

TABLE 3

Average data obtained per speech form

Speech form	T_1	T_2	Errors per questionnaire
DEC71-HO	2.87	8.29	2.0
DEC73-OO	2.71	7.20	2.17
DEC73-OE	3.41	5.31	3.33
JUN74-OE	2.85	4.30	2.0

Averages of synthetic versus natural speech data

Speech form	T_1	T_2	Errors per questionnaire
Synthetic	2.96	[illegible].27	2.37
Natural	2.97	4.52	2.29

Average data per text

Speech form	Text	T_1	T_2	Errors per questionnaire
Synthetic	A	3.05	6.17	2.5
Synthetic	B	2.87	6.38	2.25
Natural	A	3.20	5.72	2.5
Natural	B	2.78	3.31	2.08

Differences Between Particular Speech Forms

Results from the pair comparison study were analyzed and relative distances were computed on a seven-point scale. These values are plotted in Figure 1. The JUN74-OE combination (of synthesizer program and input) ranks highest with the DEC73-OE and DEC73-OO combinations occupying the next two positions, respectively, at equal intervals of about 1.6 scale points. The DEC71-HO algorithm was rated lowest--well below the other three.

Analysis of the parameter T_2 among the different synthetic speech forms does not yield sufficiently low values of probability to justify rejecting the null hypothesis, although p is always less than 0.5. However, the number of samples available in each case is very small and the variance of the measurements

LISTENERS' RELATIVE PREFERENCES AMONG DIFFERENT SPEECH FORMS

(VALUES SCALED FROM 0-7)

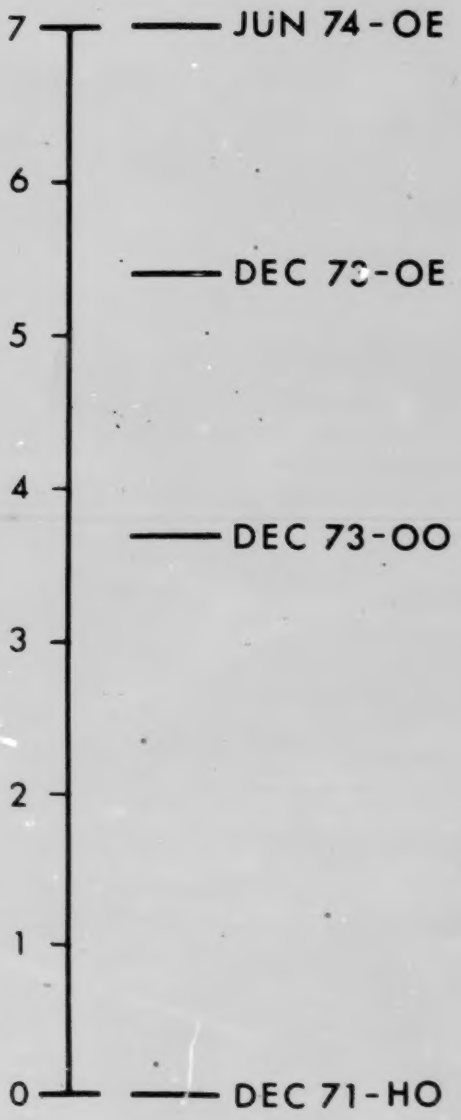


Figure 1: Preference data were obtained from 24 subjects who heard samples of the four synthetic speech forms presented in pairs. The data are ranked on an arbitrarily chosen seven-point scale.

is high owing to large individual differences among listeners. Given these circumstances, it is quite likely that more data would enable a statistical test to discriminate between each of the synthetic speech forms. Meanwhile, it is of significant interest that the average values of T_2 for the four versions of synthetic speech correlate closely with the rank order derived from the preference test. These results, plotted in Figure 2, show that the synthetic speech forms requiring the shortest period T_2 to fully complete the questionnaire are also those that are placed highest on the preference scale.

DISCUSSION

Comparing Natural Versus Synthetic Speech Comprehension

Comparison of the performances on natural and synthetic speech passages revealed a surprisingly small difference in favor of natural speech. The reasons for this finding (which was not expected on the basis of earlier intelligibility tests) may stem from some inherent characteristic of the comprehension test itself or what, in a sense, might be called "weakness" in its administration. Such possible weaknesses include the simplicity of the texts (intellectually more demanding texts might have revealed a greater difference) and the relatively slow speaking rates that were used. The question of how the subject matter affects the relative comprehension scores on synthetic and natural speech has never been systematically examined and therefore further study will be needed. Regarding the speaking rate, its effects on natural speech comprehension are well-known (Fairbanks, 1957a, 1957b), although the degree to which the observations apply to synthetic speech have yet to be ascertained. Nevertheless, setting this issue aside, there is one known consequence of speaking rate that may have specifically favored natural speech. The natural speech tape being physically shorter, could be scanned at a slightly faster rate than any of the synthetic speech tapes, and this would be expected to have a tendency to reduce the natural speech parameter T_2.

Concerning the question of speech improvement, the results in Table 3 suggest that the DEC73-OO combination of input, synthesizer, and algorithm generates better speech than the combination represented by DEC71-HO. Both algorithms received the same phonetic input derived from the stored dictionary of the Reading Machine program, but the earlier routine employs the Laboratories synthesizer while the later version uses the OVE-III.

The effects of the hand-prepared phonetic text are illustrated by the results of DEC73-OO and DEC73-OE outputs. These favor the hand-prepared texts and indicate that the linguist's knowledge of phonology, syntax, and semantics, which is brought to bear when applying adjustments, gives a measurable advantage over the computer, which applies contextual adjustments at only a very superficial level.

Finally, it is reassuring that the average times obtained on each speech form rank in a logical order--the most recent algorithms and the most carefully prepared inputs yielding the best performances. Moreover, these times agree well with the results of the pair-comparison test (see Figure 2). Taken together the data indicate that at the present stage of synthetic speech research, there is a direct relationship between listener preference and listener performance and that efforts to make the speech sound more natural (i.e., attractive) will be likely to result in significant gains in comprehensibility.

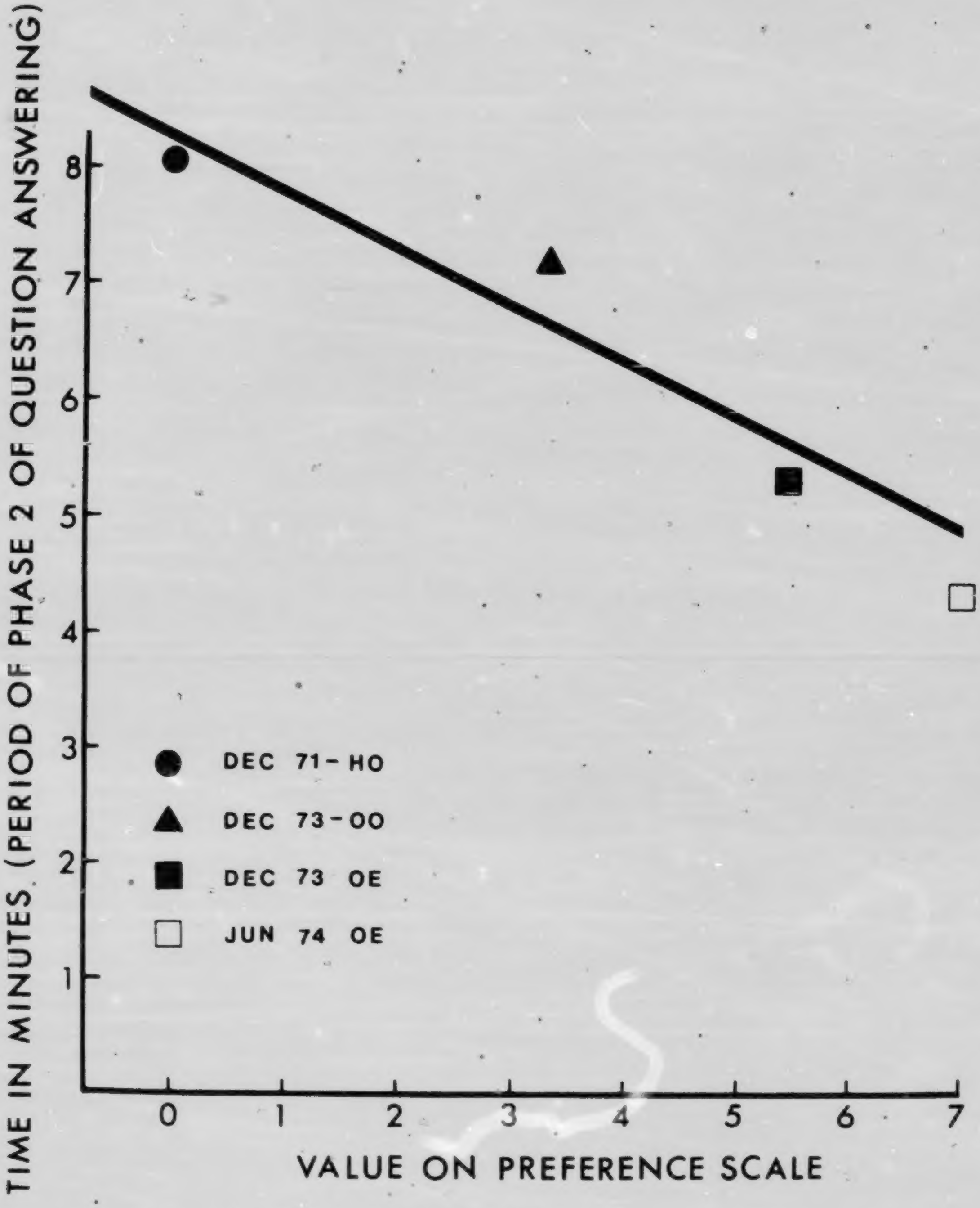


Figure 2: A plot of the time taken to complete phase 2 of the question-answering process (regarded as a measure of comprehensibility) versus the position occupied by each speech form on the preference scale.

REFERENCES

Cooper, F. S., J. H. Gaitenby, I. G. Mattingly, P. W. Nye, and G. N. Sholes. (1972) Audible outputs of a reading machine for the blind. Haskins Laboratories Status Report on Speech Research SR-29/30, 91-95.

Fairbanks, G. (1957a) Effects of time compression upon the comprehension of connected speech. J. Speech Hearing Dis. 22, 10-19.

Fairbanks, G. (1957b) Auditory comprehension of repeated high-speed messages. J. Speech Hearing Dis. 22, 20-22.

Guilford, J. P. (1954) Psychometric Methods. (New York: McGraw-Hill), chap. 7.

Ingemann, F. (1975) Testing synthesis-by-rule with the OVEBORD program. Haskins Laboratories Status Report on Speech Research (this issue).

Kuhn, G. M. (1973) A two-pass procedure for synthesis-by-rule. J. Acoust. Soc. Amer. 54, 339(A). [Also in Haskins Laboratories Status Report on Speech Research (in preparation).]

Liljencrants, J. C. W. A. (1968) The OVE-III speech synthesizer. IEEE Trans. Audio Electroacoust. AU-16, 137-140.

Mattingly, I. G. (1968) Synthesis by rule of General American English. Ph.D. dissertation, Yale University. (Issued as Supplement to Haskins Laboratories Status Report on Speech Research.)

Nye, P. W. and J. H. Gaitenby. (1973) Consonant intelligibility in synthetic speech and in a natural speech control (modified rhyme test results). Haskins Laboratories Status Report on Speech Research SR-33, 77-91.

Nye, P. W. and J. H. Gaitenby. (1974) The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. Haskins Laboratories Status Report on Speech Research SR-37/38, 169-190.

Raygor, A. L. (1970) Reading Test. [New York: McGraw-Hill Basic Skills System], pp. 7-9(A) and 6-9(B).

Testing Synthesis-by-Rule with the OVEBORD Program

Frances Ingemann*
Haskins Laboratories, New Haven, Conn.

INTRODUCTION

In 1973 a control routine for the new OVE-III synthesizer was written at the Laboratories (Kuhn, 1973). This control routine is part of a larger editorial program called OVEBORD. The synthesis subroutine converts input strings of phoneme symbols into output strings of synthesizer-parameter time frames by a two-pass algorithm. The editorial program allows the user easy on-line specification of the synthesis variables. Such user-controlled variables include the acoustic features underlying the individual phonemes as well as certain aspects of the allophone rules, which select particular variant representatives of a phoneme according to the phoneme's environment. (A description of the OVEBORD program will be found in a later issue of the Haskins Laboratories Status Report on Speech Research.)

To initialize the new program, variable-values comparable to those used with a synthesizer at the Linguistics Department of the University of Connecticut were used. The speech produced by OVEBORD with these starting values was generally agreed to sound more natural than speech on previous synthesizers at the Laboratories. Nonetheless, it was anticipated that these starting values were not optimal for the OVEBORD program either with respect to intelligibility or to naturalness.

During the early part of 1974, the present author began to work on the variable-values for OVEBORD following an approach originally attempted in 1957 (Ingemann, 1957a, 1957b). Insofar as possible, the same, or very similar, specifications are used for all members of a natural phonetic class. By mid-1974 four sets of program variable-values had been accumulated, and their performance in synthetic speech was compared by means of listening tests. The four sets of values submitted to such tests are

---

*Also Department of Linguistics, University of Kansas, Lawrence.

Acknowledgment: I want to thank Gary Kuhn for his patience in introducing me to the workings of the computer and the program, and for the modifications that he made in the program at my suggestion. I have also profited from the many conversations we have had during the course of this work.

[HASKINS LABORATORIES: Status Report on Speech Research SR-41 (1975)]

JUL73	A set of values selected by Kuhn and first made available in July 1973.
DEC73	Essentially the same set with minor modifications.
MAY74	A rather different set of values, selected by the present author according to the phonetic-class principles mentioned above.
JUN74	Modified version of the MAY74 set.

Three tests using these OVEBORD values have been completed. The first test compares the intelligibility of running synthetic speech with the DEC74 and MAY74 values. The second test compares intervocalic consonant intelligibility with the JUL73 and JUN74 values. The third test compares intelligibility of initial consonants, vowels, and final consonants with the DEC73 and JUN74 values. [The DEC73 and JUN74 values were also used in a listener comprehension and preference test reported in Nye, Ingemann, and Donald (1975).]

RUNNING SPEECH

Forty-two sentences were constructed from the 27 C-MU allophone sentences (Shockey, 1974) principally by dividing long sentences into two shorter ones. These sentences seemed appropriate for testing the rules because they had been designed to include all the sounds of English in a variety of phonetic environments. They were also particularly advantageous for us since many were, as the author of the C-MU sentences expressed it, "weird in lexical content"; consequently, phoneme recognition was more crucial than it might be in more predictable sentences. The sentences contained 347 words and the phonemic input for synthesis consisted of 1070 segmental units.

The 42 sentences were synthesized using the DEC73 and MAY74 rules and were also read by a human speaker. The sentences were divided into 3 sets of 14 sentences each. Subjects heard one set of the natural speech and one set of each of the two synthetic versions. No subject heard more than one version of any sentence. Twenty-four subjects in all participated in the experiment so that each version of each set was heard by eight subjects. The subjects were all people associated with the Laboratories, most of whom had previous exposure to synthetic speech.

Each sentence was spoken twice and subjects were asked to write the sentence they heard. The results were

	DEC74	MAY74	Natural
Words correct (347 tokens)	71%	68%	99%
Words correct (235 types)	68	65	99
Phonemes correct (1070 tokens)	78	75	99

Percentages of correct identifications of individual phonemes are given in Table 1. In assessing these scores, it should be noted that many words, phrases, and sometimes even whole sentences were omitted in the subjects' responses. Undoubtedly, subjects recognized some of the sounds intended in these omitted portions even though they did not recognize enough to enable them to write anything meaningful.

TABLE 1: Analysis by phoneme of correct responses to the C-MU sentences listening test.

Phoneme	Number of Occurrences	Percent Correct DEC 73	 MAY 7
ʃ	4	78	91
o	13	91	89
g	14	77	88
s	43	84	84
ɝ	22	85	84
f	27	85	83
w	22	80	83
ʊ	7	77	82
ɔ	16	84	82
ɛ	21	81	82
a	21	78	81
y	12	77	80
au	10	93	80
ə	91	85	80
p	29	78	78
aɪ	20	86	78
i	34	82	78
l	40	81	78
k	33	82	77
e	17	75	76
æ	32	82	75
ʌ	13	75	75
ɪ	68	71	74
r	46	81	74
ð	46	73	73
m	31	88	73
b	25	83	72
ɔɪ	6	60	71
n	61	81	70
h	16	71	69
z	39	70	68
j	9	61	68
t	87	73	68
d	27	64	66
u	20	75	64
θ	11	65	63
v	19	63	62
ŋ	10	65	60
ɔr	3	71	58
č	3	41	50
ʒ	2	25	31
Totals	1070	78%	75%

Because of a defect of the testing procedure, the results may have been slightly poorer than they would otherwise have been. When the first few subjects were run, insufficient time was allowed to write the sentences comfortably. As a result, errors and omissions may have occurred when a subject had not finished writing one sentence before he heard the next. Since the speaking rate for the MAY74 rules was slightly faster and the interval between stimuli slightly shorter, more such errors may have been made for MAY74 rules than for DEC73 rules. Another listening test correcting these defects is planned for these same sentences using the DEC73 and JUN74 rules.

An inspection of the places where subjects made errors suggested some changes that could be made in the MAY74 rules. These revised rules JUN74 were used in the other two listening experiments.

INTERVOCALIC CONSONANTS

In July 1973 Kuhn conducted a test of intervocalic consonants in which each of the 24 consonants of English occurred once in each of the following environments:

i___i	a___i	u___i
i___a	a___a	u___a
i___u	a___u	u___u

The resulting 216 stimuli (each played twice) were randomized and presented to six listeners who were asked to identify the consonants.

For purposes of comparison, the same vowel-consonant-vowel (VCV) sequences in the same order were synthesized using the JUN74 rules and presented to the same six subjects 11 months later. The results were

	JUL73	JUN74
Percent correct	74	80

A confusion matrix of the responses to the JUN74 rules is given in Table 2.

CONSONANT-VOWEL-CONSONANT (CVC) UTTERANCES

Four lists of 50 monosyllabic words each devised by Mitchell (1974) to test 22 consonants in initial position, 13 consonants in final position, and 15 vowels and diphthongs in medial position were used to test DEC73 and JUN74 rules. For each stimulus, five alternative responses are provided that differ in one phonetic feature at a time. Mitchell had found these lists to be 98 percent intelligible to listeners with normal hearing. Since these lists were developed for clinical use with hard-of-hearing listeners, they did not always provide responses suitable for confusions that occur in listening to synthetic speech. Therefore, listeners to the synthetic versions were allowed to write in what they heard if it was not one of the words provided. These write-in responses were scored correct if the particular phoneme being tested was correctly identified; for example, *lib* instead of *lip* was considered a correct response to a stimulus intended to test initial *l*.

TABLE 2: Responses to the intervocalic consonant listening test, synthesized by JUN74 rules.

TOTAL RESPONSES FOR EACH PHONEME

(6 subjects x 9 stimuli = 54)

RESPONSE

PHONEME INTENDED	b	p	m	w	d	t	n	l	g	k	ŋ	r	ǰ	č	y	v	f	ð	θ	z	s	ʒ	ʃ	h	ø	Other
b	47	7																								
p		54																								
m	5		36						1		5															7
w				46				3				2				2								1		
d					43				10									1								
t		2				50				1																1
n							29	3	2		14															6
l								54																		
g					11				43																	
k		1				18			1	34																
ŋ								1	10		31				4											8
r												54														
ǰ									1				53													
č														54												
y								1							52										1	
v	2		1	4				2	4			20				20		1								
f																1	53									
ð					1			8				1			24			12	2					5		1
θ					2										1			6	27	13	2			2	1	
z															1					49	4					
s																					54					
ʒ													9									37	8			
ʃ																							54			
h									1															48	2	3

TABLE 3: Responses to stimuli testing initial consonants on the Mitchell lists.

RESPONSES

SYNTHESIS

DEC 73

	b	p	m	w	d	t	n	l	g	k	ŋ	r	ǰ	č	y	v	f	ð	θ	z	s	ʒ	ʃ	h	ø	Other
b	10	1	1													4										
p	1	14															1									
m	5		10	1																						
w				16																						
d	1				7		5											3								
t					2	12				1															1	
n							11	4							1											
l								16																		
g									16																	
k						3			3	10																
ŋ																										
r												16														
ǰ					1								15													
č														16												
y															16											
v	1			1								2				12										
f																	16									
ð					1			1							2			12								
θ																			12		2	1		1		
z								1												14						1
s																					16					
ʒ																										
ʃ													1	4									11			
h										3														12	1	

JUN 74

	b	p	m	w	d	t	n	l	g	k	ŋ	r	ǰ	č	y	v	f	ð	θ	z	s	ʒ	ʃ	h	ø	Other
b	11															5										
p	2	12	1																					1		
m	6		10																							
w				16																						
d					12													3							1	
t					2	7																		6	1	
n			3		1		10	2																		
l								15																	1	
g	1				2				10	1		1													1	
k		1	1						2	11														2		
ŋ																										
r												16														
ǰ													15		1											
č													7	9												
y															15										1	
v	1		1	1								2				11										
f																	16									
ð					1											3		12								
θ																			10	2	4					
z					1															10	4					1
s																					16					
ʒ																										
ʃ																			1	1	3		11			
h						3				11									1					1		

TABLE 4: Responses to stimuli testing final consonants on the Mitchell lists.

RESPONSES

SYNTHESIS JUN 74	b	p	m	d	t	n	g	k	f	v	ð	z	s	ʃ	θ
b	12	3	1												
p	1	14			1										
m	1		11			3									1
d				12	3		1								
t					15			1							
n						16									
g							15	1							
k				2	3		5	6							
f									12	4					
v	2		1							10	2				1
ð				1		1				2	11				1
z						1						15			
s													16		

SYNTHESIS DEC 73	b	p	m	d	t	n	g	k	f	v	ð	z	s	ʃ	θ
b	16														
p	8	7		1											
m	4		9			2				1					
d				13	2		1								
t				3	11			2							
n			1			15									
g				1			15								
k				1			4	11							
f									12			1	1	1	1
v			1							11	2	1			1
ð				2		1				2	9	1			1
z											1	15			
s												1	15		

TABLE 5: Responses to stimuli testing vowels and diphthongs on the Mitchell lists.

RESPONSES

SYNTHESIS

JUN 74

	i	ɪ	e	ɛ	æ	ɑ	ɔ	o	ʊ	u	ʌ	ɝ	ɑɪ	ɔɪ	ɑʊ	∅
i	16															
ɪ		16														
e			16													
ɛ		1		15												
æ				1	15											
ɑ						16										
ɔ							15				1					
o								16								
ʊ									16							
u										16						
ʌ											16					
ɝ												16				
ɑɪ			1										15			
ɔɪ													1	15		
ɑʊ															16	

DEC 73

	i	ɪ	e	ɛ	æ	ɑ	ɔ	o	ʊ	u	ʌ	ɝ	ɑɪ	ɔɪ	ɑʊ	∅
i	16															
ɪ		16														
e			16													
ɛ				16												
æ				1	15											
ɑ						12					4					
ɔ							16									
o								16								
ʊ									16							
u										16						
ʌ							2				14					
ɝ												16				
ɑɪ													16			
ɔɪ														16		
ɑʊ															16	

Listeners sometimes need a brief exposure to synthetic speech before they begin to hear it as speech. So each list was preceded by the following instructions synthesized by the same rules that were to be tested.

> You will hear a number followed by one of the five choices listed on the answer sheet. The word will be said only once. Please circle the word you hear. If you hear none of the words on the answer sheet, you may write what you do hear in the right-hand margin.

The instructions were also printed on the response sheet.

Each listener heard one list synthesized by the JUN74 rules and another by the DEC73 rules. Since four listeners heard each list and each phoneme to be tested occurred once in each of the four lists, there were a total of 16 judgments for each phoneme in each synthesis version. The results were

	DEC73	JUN74
22 initial consonants	73%	82%
13 final consonants	76	79
15 vowels and diphthongs	97	98
Total	81%	86%

Confusion matrices are given in Tables 3-5.

CONCLUSIONS

The various sets of variable-values tested do not differ greatly in the intelligibility of the speech they generate. If any set has an edge on the other, it is probably the JUN74. From this it seems safe to conclude that using similar values across natural phonetic classes causes no serious deterioration in the synthetic speech. It is also apparent from these tests that this synthetic speech does not yet approach the intelligibility of natural speech.

For reasons of clarity, scores have been presented by individual phonemes in the various tests. This is an oversimplification and it should not be allowed to obscure the fact that some sounds are highly identifiable in certain environments and poorly identifiable in other environments. Future improvement of the variable-values should result from systematic investigation of these poorer sounds according to the specific contexts in which listeners have difficulty in identifying them.

REFERENCES

Ingemann, F. (1957a) Speech synthesis by rule. J. Acoust. Soc. Amer. 29, 1255(A).

Ingemann, F. (1957b) Rules for synthesizing American English on the pattern playback. Unpublished manuscript.

Kuhn, G. M. (1973) A two-pass procedure for synthesis by rule. J. Acoust. Soc. Amer. 54, 339(A).

Mitchell, P. D. (1974) Test of differentiation of phonemic feature contrasts. J. Acoust. Soc. Amer., Suppl. 55, S55(A).

Nye, P. W., F. Ingemann, and L. Donald. (1975) Synthetic speech comprehension: A comparison of listener performances with and preference among different speech forms. Haskins Laboratories Status Report on Speech Research SR-41 (this issue).

Shockey, L. (1974) Description of C-MU allophone sentences. ARPA Network Information Center, SUR Note 128.

Stress and the Elastic Syllable: An Acoustic Method for Delineating Lexical Stress Patterns in Connected Speech*

Jane H. Gaitenby
Haskins Laboratories, New Haven, Conn.

ABSTRACT

Particular lexical stress patterns are common to the speech of native talkers of standard American English, but these patterns may be produced in a variety of prosodic ways. In this report, a technique is described for the retrieval of prosodic contours from the acoustic record (of a sentence as read separately by four individuals) that agree well with lexical stress patterns, as perceived in a pilot study.

INTRODUCTION

The main purpose of this report is to describe a method of deriving lexical stress patterns from the acoustic record of connected speech. Secondarily, it will be suggested that larger prosodic contours may be revealed by the same method. Prosodic data for one long sentence will be presented, and will be compared by talker, by prosodic parameter, and by summed parameter values in sequential syllables.

In this report, stress is defined as the property that endows sequential syllables with differentiating grades of acoustic prominence. The prosodic features that interact in signaling stress are: fundamental frequency--to be referred to below, with prosodic license, as "pitch"--duration, and intensity. Selected acoustic measurements of these three features comprise the data for the study. Spectral distribution, which is also generally acknowledged to be a stress cue, will not be explicitly referred to here.

The state of stress research can be summarized by noting that a great deal of what is known, and of what is known to be unknown, on the subject of stress

---

*This report is an expanded version of "The Elastic Syllable: An Acoustic View of the Stress-Intonation Link," a paper that was presented at the 88th meeting of the Acoustical Society of America, St. Louis, Mo., 4-8 November 1974. [J. Acoust. Soc. Amer. (1974), Suppl., 56, S32 (Abstract P5).]

Acknowledgment: Franklin S. Cooper introduced the author to the investigation of prosodic problems in English speech, such as the one described, and has provided guidance at just the right intervals. This is deeply appreciated. Warm thanks, too, to John M. Borst for his patient advice on instrumentation and measurement.

today, was well described--under the heading of "Accent"--as early as 1934 by Carhart and Kenyon (1934) in the Guide to Pronunciation in the second edition of Webster's New International Dictionary.

[For further general background on stress research, the reader is referred to a concise review of the literature given by McClean and Tiffany (1973) in introductory paragraphs to their article. Also provided in the article are significant acoustic data and observations on effects of position, loudness, and rate on stress realization.]

Stress research is complicated by the fact that the three acoustic parameters acknowledged as cosignals to stress also apparently share in signaling another speech attribute, namely, "intonation." Intonation is thought by many to refer only to the perceptual phenomenon of pitch variation across an utterance, and the majority of intonation studies accordingly have been concentrated only on fundamental frequency contours. [Noteworthy exceptions are Denes (1959), Denes and Milton-Williams (1962), and Lieberman (1967).]

A further problem in investigating stress is that there are several types of stress that should be distinguished: lexical, semantic (under which we include contrastive and emphatic stress), and positional stress. These may cooccur in speech and thus confound analysis.

Another fact that makes stress description and analysis difficult is that stress perception is dynamic (as indeed is all speech perception), but acoustic displays of speech, such as spectrograms, immobilize the speech wave, leading to descriptions of the physical record that appear to deal with static events.

Published objective descriptions of stress have been fragmentary. If the corpus of speech examined in a study is relatively long, then the stress-signaling parameters described are probably few. Conversely, if two or three prosodic parameters are dealt with in detail, then the corpus itself is probably brief--consisting of nonsense syllables, single words, or extremely short sentences. (Furthermore, spontaneous natural speech is seldom used in stress experiments; text readings are preferred because they provide controlled verbal content.) In physiological research on stress, for which improved instrumentation and analytic techniques have been arduously developed in recent years, published reports have thus far, understandably, been confined to the behavior of only scattered portions of the vocal apparatus. Finally, very few accounts of stress experiments involve parallel data from more than one or two of the possible approaches to speech research, which may be physiological, acoustic, perceptual, and synthetic. [Lieberman (1967) is one of the exceptions.] Therefore, the analysis of stress remains partial and primitive, owing to the lack of multifaceted data on sizable stretches of natural connected speech.

[It may seem somewhat surprising that speech synthesis by rule is as good as it is, in view of the poverty of information available on stress. One reason for the high intelligibility of some versions of current synthetic speech must lie in the adequacy of the segmental rules used, including extremely good rules for duration. (Duration is the most tightly structured of the prosodic features in English, as will be illustrated below in natural speech data.) Aside from duration, considerable prosodic variation (elasticity within and across syllables) is permissible in the language. Mild prosodic (and phonetic) deviations from the norm, such as those heard in synthetic speech, may be heard as dialects--to which most listeners can adjust themselves, as long as the variations are regular.]

Having expressed the need for more extensive investigations of stress, and having produced and noted the preceding caveats, the scope of this paper is nevertheless limited, in that it deals only with data from the acoustic plane. In this paper we describe an approach to the characterization of lexical stress by way of a measurement and display technique that delineates acoustic patterns corresponding closely to intrinsic stress patterns.

The present paper offers a new look at prosodic measurements that we made between 1958 and 1960. The material pertaining to the speech sample used, the method of measurement, and the measurement units themselves therefore date from that time, when the purposes of the experiment were to produce an acoustic description of running speech and to find correlates of stress in that acoustic record. It was thought sufficient at that time to characterize acoustic stress in a relative manner, by merely noting whether the combination of pitch, duration, and intensity parameters in a syllable were higher or lower than the prosodic combinations in immediately adjacent syllables. In recent reexaminations of the same acoustic data, it has appeared that more informative stress patterns can be revealed by referring to the absolute prosodic measurements. It is this latter approach that will be presented, after the procedure used in the initial acquisition of the data has been described.

I. DATA ACQUISITION

A. Initial Assumptions

Two assumptions were made at the outset of the original experiment:

1. Peak pitch, peak intensity, and total duration of voicing in a syllable are sufficient data to characterize syllable stress (relative to adjacent syllables).

2. Syllabic acoustic data for these three prosodic parameters can be combined to produce total prosodic (stress) value of a syllable. (The three parameters share the attribute of signaling stress perceptually; it is therefore reasonable to assume that acoustic parameters combine to signal stress.)

B. The Corpus

The speech material used consisted of readings of a text (about 500 words long) that was created from a selection of high-frequency English vocabulary, including polysyllables as well as monosyllables (Dewey, 1923; Thorndike and Lorge, 1944). Several of the polysyllables were intentionally repeated, at least once in the script, in contrasting locations, and in differing grammatical roles where that was possible, e.g., "official" was used as a noun in one sentence, and as an adjective in another. Words in which stress patterns change with grammatical, semantic, or positional usage (such as "transport," "invalid," "absolute") were not used.

The form and content of the text was like a dull governmental announcement (high-frequency polysyllables from word counts of printed matter suggest that semantic field) and most of the sentences were long. Unemphatic readings at normally fast speaking rates were required, and it was assumed that the long and uninteresting sentences would contribute to those effects. It was also anticipated that the intrinsic stress patterns of the polysyllables would be very reduced in such a context; therefore, any evidence of acoustic correlations with lexical stress patterns might be considered basic stress cues.

The text was read, casually and rapidly, by three men and one woman from the laboratory staff. Each person was recorded at a tape speed of 15 ips under standard sound-proofed room conditions. The talkers were native to the United States and spoke "eastern educated speech," although their maturational years were spent in various parts of the country. Their ages ranged from 30 to 42 years.

Three focal sentences, containing among them two or more instances of certain polysyllables, were excised from each person's tape recording and were then measured by the means to be described. The shortest of the sentences (29 syllables long) will be discussed in this paper.

C. Measurement Method

The speech waveform and hill-and-dale trace of the pitch level were recorded by dual-beam cathode ray tube on 35-mm film at 7.2 ips. (The pitch voltages were taken from a conventional Vocoder.) The pitch values were calibrated against 125-msec tape-recorded sequences of 80-, 90-, and 120-Hz pure tones that had been spliced into the source audio tapes. Measurements of peak pitch and total duration of voicing were made by reference both to the film and to wide and narrow band spectrograms; the amplitude curves above the wide band displays were used for the peak intensity measurements.

The syllable boundaries were marked consistently and in corresponding positions on the film and spectrograms, for the four versions of the sentence, with word boundaries preserved because word stress patterns were to be compared.

D. Measurement Units

It will be seen (Figure 1) that the units of measurement are not conventional, although the pitch and duration data shown can be converted readily to traditional units, as will be described. The intensity data shown will also be explained.

Three constraints were taken into consideration before the decision was made on how the acoustic measurements might best be examined and presented as prosodic data:

1. There were limitations on the precision of measurement, imposed by the small size of the acoustic displays employed. For example, the resolution of the pitch trace (on film) permitted measurements no finer than in approximately 4-Hz units.

2. The numerical ranges of all three parameters had to be compatible in magnitude so that the parameters could be displayed and compared in parallel on a common grid.

3. Weighting of the parameters seemed desirable. [Bolinger (1958) had presented evidence for the primacy of pitch over duration as a

stress cue, and Fry (1958) had shown that duration was a stronger stress cue than intensity (in single words, at least).] It seemed reasonable, then, to approximate the apparent hierarchy of cues in the graphic display of the prosodic measurements.

A practical method of working within these constraints was to estimate the likely ranges of the three parameters to be found in the four readings, and then to scale the separate parameters to represent the stress cue primacy of pitch over duration, and duration over intensity. To do this, acoustic measurements were converted to representative "prosodic units."

For pitch, the lower limit of the range was set at 60 Hz. [All syllables in which the pitch peak registered 60 Hz or below are called "0" (zero) in the Figure 1 data because very low pitch values are usually accompanied by low intensity levels, which are normally below the threshold of hearing.] An upper limit of about 200 Hz was assumed on the basis of preliminary inspections of the acoustic record. The resulting range of 140 Hz (60-200 Hz) was measured in 4-Hz units, producing a range of 35 "prosodic units." A range of 35 steps seemed sufficient for the display of syllable peak pitch contours.

The durational range was estimated at 20 to 400 msec, for the shortest to the longest syllables, voiced portions only. It was appropriate to measure duration in 20-msec units, which produced 20 prosodic units as equivalents to the anticipated durational range.

Syllable peak intensity measurement was made from a logarithmic plot of the rms voice amplitude in dB (the amplitude curve displayed on the Kay Sonagraph spectrogram). Maximum and minimum intensity values were found for each talker (i.e., a personal vocal intensity range). This range was divided into 14 equal linear steps--the smallest practical number of divisions. These were called the 14 prosodic units of intensity. Consequently, the prosodic unit of intensity may differ somewhat from talker to talker, but it is consistent across the utterance for each individual.

In short, to make weighted graphic comparisons, we measured the parameters in prosodic units that represented actual measurements on each parameter, but the number of prosodic units available to the display of each separate parameter was apportioned to suggest the rank ordering of the stress cues. Thus:

Parameter	Range of Prosodic Units
Pitch	35
Duration	20
Intensity	14

It must be emphasized that the values of the prosodic units for different parameters have been selected both as a weighting device and for graphic convenience. Actual stress equivalence is NOT implied between, for example, 10 prosodic units of pitch and 10 prosodic units of duration or intensity, although, for the purpose of the analysis to follow, they will be treated as equivalent.

SYLLABLE PEAK PITCH, DURATION OF VOICING, AND PEAK INTENSITY IN SELECTED ABSOLUTE UNITS; DATA FOR FOUR TALKERS

Sentence: An official, that is a department head, hopes that you will understand what several of the officers' comments mean.

		An	o	ffi	cial,	that	is	a	de	part	ment	head,	hopes	that	you	will	un	der	stand	what	se	veral	of	the	o	ffi	cers'	co	mments	mean.
TALKER R	Pitch	12	10	19	13*	13	13	10	9	11	8	8*	15	6	12	11	13	10	11	8	13	10	8	9	13	0	10	11	11	8
	Duration	5	2	4	6*	11	7	2	5	6	6	9*	3	5	5	5	3	4	18	6	5	6	5	5	9	2	5	5	10	14
	Intensity	10	10	11	10*	12	12	11	10	12	11	11*	11	9	11	11	12	13	11	11	12	11	9	10	10	8	10	12	11	12
	TOTAL	27	22	34	29*	36	32	23	24	29	25	28*	29	20	28	27	30	27	40	25	30	27	22	24	32	10	25	28	32	34
TALKER S	Pitch	12	8	15	0*	8	8	0	5	0	0	7*	18	1	15	15	14	13	13	14	15	7	12	0	10	0	0	12	0	2
	Duration	4	3	4	6*	6	6	2	1	6	3	9*	5	6	3	6	7	4	10	3	5	3	4	5	5	2	3	5	9	12
	Intensity	10	10	11	5*	9	8	5	4	6	5	5*	8	6	9	10	11	10	9	7	11	8	8	8	8	2	3	8	9	8
	TOTAL	26	21	30	11*	23	22	7	10	12	8	21*	31	13	27	31	32	27	32	24	31	18	24	13	23	4	6	25	18	22
TALKER G	Pitch	9	15	16	15*	16	19	0	17	18	15	18*	14	0	10	15	11	13	18*	18	12	2	7	17	12	0	0	13	16	18
	Duration	4	3	4	10*	7	15	4	4	6	6	17*	6	2	5	6	7	4	24*	4	5	11	6	5	7	2	8	4	14	19
	Intensity	8	9	10	8*	10	11	3	7	9	8	9*	9	5	5	7	6	6	9*	7	7	6	6	6	5	4	5	4	10	9
	TOTAL	21	27	30	33	33	45	7	28	33	29	44*	29	7	20	28	24	23	51*	29	24	19	19	28	24	6	13	21	40	46
TALKER L	Pitch	11	8	18	13*	6	7	7	6	9	7	11*	16	15	14	11	11	11	9	8	11	11	8	10	8	5	8	8	5	9
	Duration	4	3	7	10*	9	10	3	3	7	5	13*	7	3	4	7	6	4	16	3	5	7	5	4	7	2	3	7	9	13
	Intensity	10	9	11	7*	8	7	7	6	9	4	5*	12	8	9	10	9	9	8	9	10	9	6	5	7	3	5	5	4	7
	TOTAL	25	20	36	30*	23	24	17	15	25	16	29*	35	26	27	28	26	24	33	20	26	27	19	19	22	10	16	20	18	29

* = PAUSE

FIGURE 1

E. Data

In Figure 1 the acoustic data in prosodic units are presented for each successive syllable of the sentence, "An official, that is a department head, hopes that you will understand what several of the officers' comments mean." The prosodic units shown there can be converted, if desired, back to traditional measurement units as follows.

For pitch in Hz, multiply the number of prosodic units given for a syllable on a pitch row by 4, and add 60. [For instance, for Talker R, first syllable, Pitch = 12. (12 X 4) + 60 = 108 Hz.]

For duration in msec, multiply the prosodic units given on a duration row for a syllable by 20.

The intensity units shown are relative within the speech of each particular talker. (The highest intensity measurement possible in any of the sentences was 14 prosodic units. In this sentence, the highest intensity value happens to be 12.)

When inspecting the data, the reader must bear in mind that the measurements refer to peak pitch in each syllable, to total duration of syllable voicing (which includes voicing in consonants as well as in vowels), and peak intensity in the syllable. The row of syllable Total values will be referred to in Figures 4 and 5 and can be ignored for the present.

II. ANALYSIS

A. Single Parameters, Compared

The data presented in Figure 1 are exploited in various graphic ways in Figures 2-5. In Figure 2, the prosodic unit data for the last portion of the sentence ("...hopes that you will understand what several of the officers' comments mean.") are shown by individual talker. The syllabic data nodes for each parameter, indicated by small circles, have been connected by (distinctive) lines in order to produce comparable prosodic contours across the utterance. Individual speaker differences in contour shapes and ranges of the trio of parameters are immediately visible. There are also resemblances across the speakers, notably in duration, as was expected (Gaitenby, 1965).

It can also be seen that the pitch contour is closely paralleled by the intensity contour (or vice versa) in the records of Talkers R and L, but these contours are less clearly related for Talkers S and G. There are other differences to note: Talker R's intensity range is quite narrow, and R also tends to use more or less equal durations in some sequences. Talker S shows considerable fluctuation in all parameters. Talker G (the female in the group, with atypically low vocal register for a female) produces pitch excursions that are more extreme, and generally higher, than those of the other talkers. G's prepausal durations are also the longest. In Talker L's record, in contrast to the others, a clear falling trend can be seen in the pitch and intensity contours (with final higher peaks).

The main point that Figure 2 is intended to illustrate is that readings of the same verbal material by several speakers produce prosodic trio contours for each talker that look far from identical from talker to talker.

PITCH vs DURATION vs INTENSITY DATA, BY INDIVIDUAL TALKER
(LAST 18 SYLLABLES OF SENTENCE)

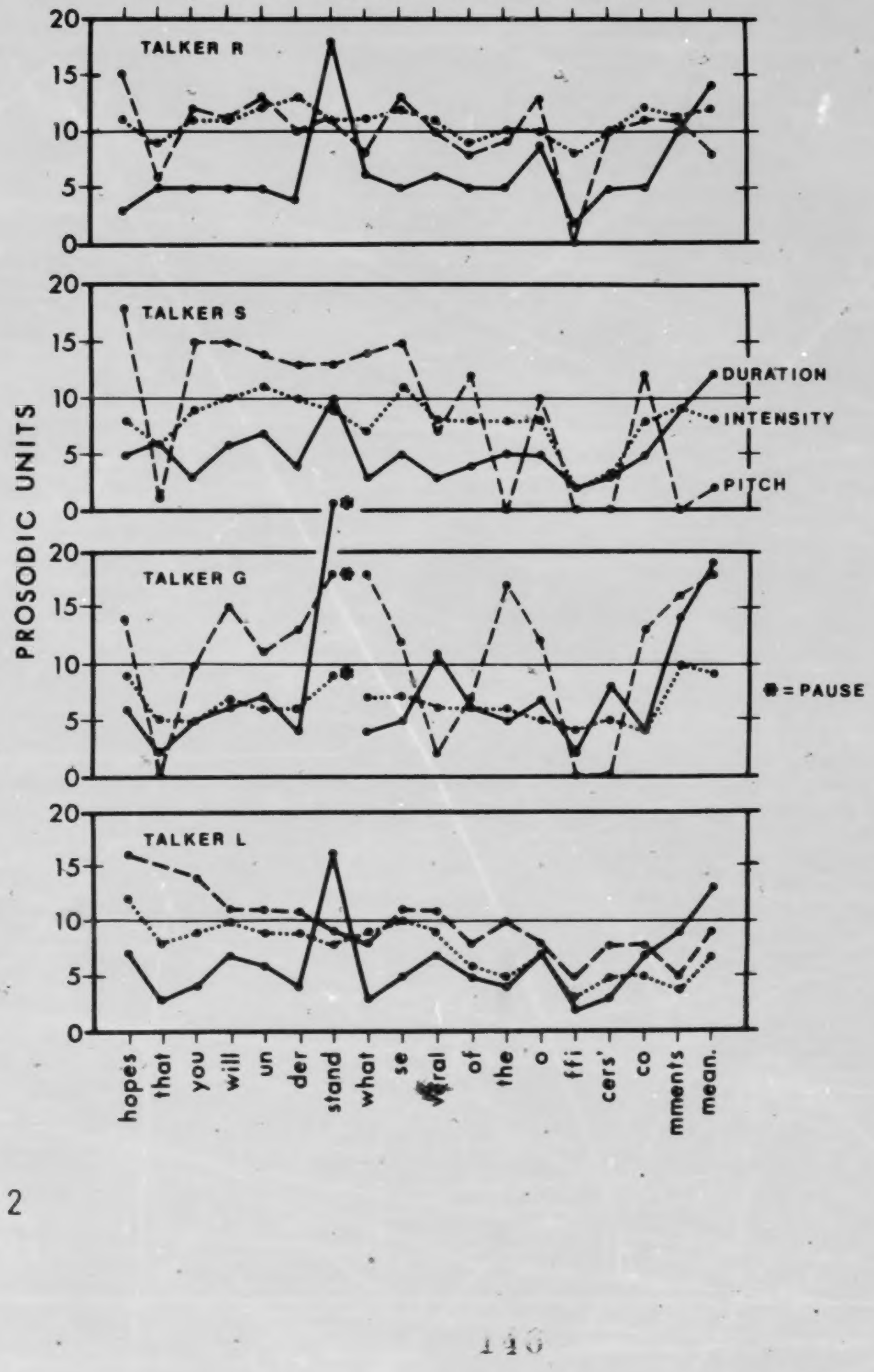


FIGURE 2

In Figure 3, cross-speaker comparisons are made by *parameter* for all 29 syllables of the sentence. We now focus on the generalized contours produced by the four talkers' versions of pitch, duration, and intensity. The talkers therefore have not been identified in these graphs (but they can be, by reference to the preceding figures).

Extreme similarity in relative duration is now obvious; there are few deviations from the essentially single pattern produced by the four individual duration contours. The talkers plainly conform in temporal organization of the sentence, although their individual speaking rates may differ. [Fundamental determinants of relative syllable duration are the number and kind of phonemes in the syllables, determined by the particular vocabulary used in an utterance. One might then assume that the common verbal material accounts entirely for the profound cross-speaker durational regularities, but preliminary studies we have made (unpublished) indicate that non-native talkers of American English produce different durational patterns from those of natives (unless their English rhythm is so "good" that it cannot be distinguished from that of native speakers). The cross-language durational problem also involves the consideration of languages in which there are phonemic length contrasts (see Peterson and Lehiste, 1960; Lehiste, 1970) and native versus non-native stress realization. However, we shall not pursue the matter here.]

Definite similarities in the four intensity contours also appear in Figure 3, as well as some general agreement (very strong in the initial phrases despite later occasionally contradictory slopes) in the overall pitch pattern. Note the four distinct pitch registers visible in the syllables of the appositive phrase, "...that is a department head,...."

The contours of all three parameters show fairly good agreement in rising and falling with the lexical stress patterns of the polysyllabic words, "official," "department," and "officers." A possible exception, at first glance, is the inherently low-stressed final syllable of "official" [l̩], in which the duration of voicing rises above that of the preceding syllable [I], seeming to indicate increased stress and/or different phonetic content. It is clear that [I] and [l̩] are different phonetically, and that [I] is intrinsically brief. However, [l̩] is prepausal, phrase final, and precedes an embedded parenthetical clause, all of which necessarily involve extended duration (Klatt, 1974). The slight rise in duration in this case is thus, at least in part, a conditioned effect. (Unless a rise in prepausal duration is very substantial--on the order of more than twice the normal length of a syllable of the given phonetic type in prepausal position--a stress increase is probably not indicated by a rise in duration alone.) Note here that all talkers' pitch *and* intensity peaks fall in [l̩], counteracting the rise in duration.

One syllable in which greatly extended duration does appear to be the major signal to increased stress is the final syllable of "understand" [ænd]. This is an intrinsically long syllable in number of voiced phonemes, and it is also phrase final. Talker G paused after this syllable; the other talkers did not produce silence here. At least two of the other talkers, however, also appear to have used length as the prime cue to the stress rise, and they may have simultaneously produced a "pseudo-pause" [a term and concept from Coker, Umeda, and Bowman (1973)] at that syntactic break by means of the prolonged syllable duration.

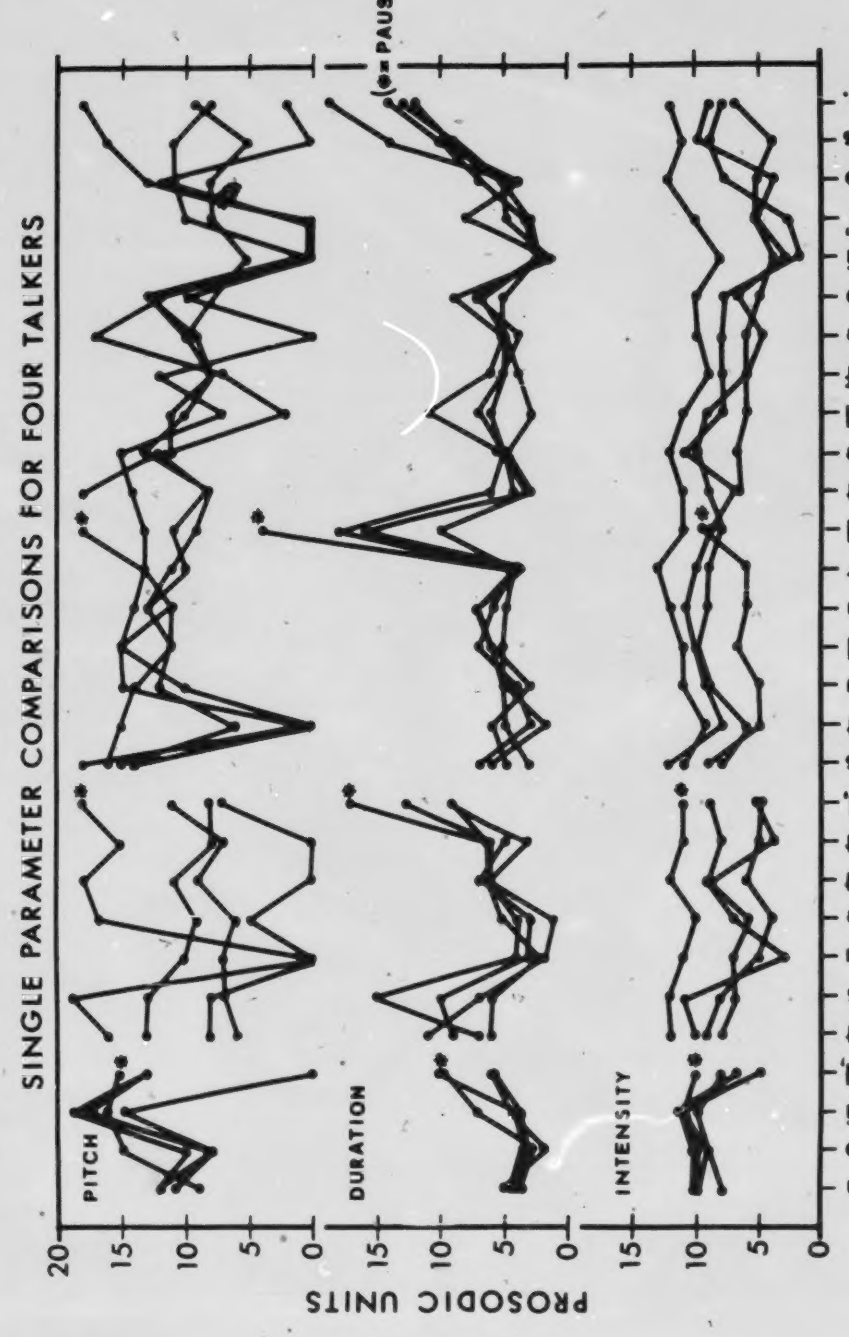

SINGLE PARAMETER COMPARISONS FOR FOUR TALKERS
PITCH
DURATION
INTENSITY
(*= PAUSE)
PROSODIC UNITS
20
15
10
5
0
15
10
5
0
15
10
5
0
An
o
ffi
cial,
that
is
a
de
part
ment
head,
hopes
that
you
will
un
der
stand
what
se
veral
of
the
o
ffi
cers'
co
mments
mean.

FIGURE 3

Note in Figure 3 that no single prosodic parameter is a reliable acoustic cue to lexical stress for all of the four talkers. That is to say, no one parameter shows unanimous rises and falls corresponding to the inherent stress patterns of the words of more than one syllable.

B. Summed Parameters, Compared

It is widely known that perceived stress depends on the combined effects of the prosodic properties within a syllable (in relation to those of adjacent syllables). It is therefore reasonable, on the acoustic plane, to sum the values of the separate parameters within each syllable, and to compare and examine the prosodic contours so produced. Summing the values of the trio of parameters is permissible because the measurements have been converted to common prosodic units.

Figure 4 shows the contours resulting from plots of the parameter totals. (The data points shown here were taken from the rows called "Total" in Figure 1.) Despite the individual differences that showed up in the single parameter contours in Figures 2 and 3, here the four talkers' separate versions of the sentence are generally alike in overall prosodic pattern. Although minor conflicts among the talkers in syllable slope are observable in these contours, most conflicts are probably due to slight differences in the semantic-syntactic interpretation of the verbal material and to idiolectal variations. Note that the lexical stresses of the polysyllabic words, however, are reflected by all of the talkers by appropriate rises and falls--in all but four instances (of which one, the [l̩] in "official," has been discussed). In the second syllable of "several" (pronounced [vrl̩] by all of the talkers), one person produced a very small contour rise (contrary to the expected pattern of the word), and two of the four talkers each produced a decided rise on the second syllable of "comments." These conflicting slopes may be artifacts of the syllabification procedure used in these words, i.e., after the first vowel. If the syllabification had followed the articulation and perception more realistically, some portion of the voiced consonant after each vowel of the first syllable would have been included in the respective first syllable of each word, thus altering at least the value of the first syllable's duration in an upward direction. (This does not explain, it must be admitted, why the other two talkers nevertheless produced the "correct" falling slopes for the second syllable.) In the case of "comments," it may be significant to the contour that the vowel in the lexically less-stressed second syllable is relatively full-grade, and furthermore, that this less-stressed syllable is penultimate in the utterance, which is to say that it is likely to have received conditioned lengthening in that location.

There are other possible explanations for the rises found in the slopes of some of the unstressed syllables. The most obvious is that vowel color was (intentionally) omitted as a prosodic parameter in this experiment, although it is a fact that no English syllable containing either a schwa or a syllabic consonant is lexically stressed. A consideration of vowel color would thus mark the syllabic "l̩" syllables as unstressed.

Another reason for the occasional slope discrepancies may lie in the weighting method itself, which can easily be modified. In particular, the conditioned effect of position on syllable duration, alluded to previously, might be compensated for in a revision of the parameter weighting.

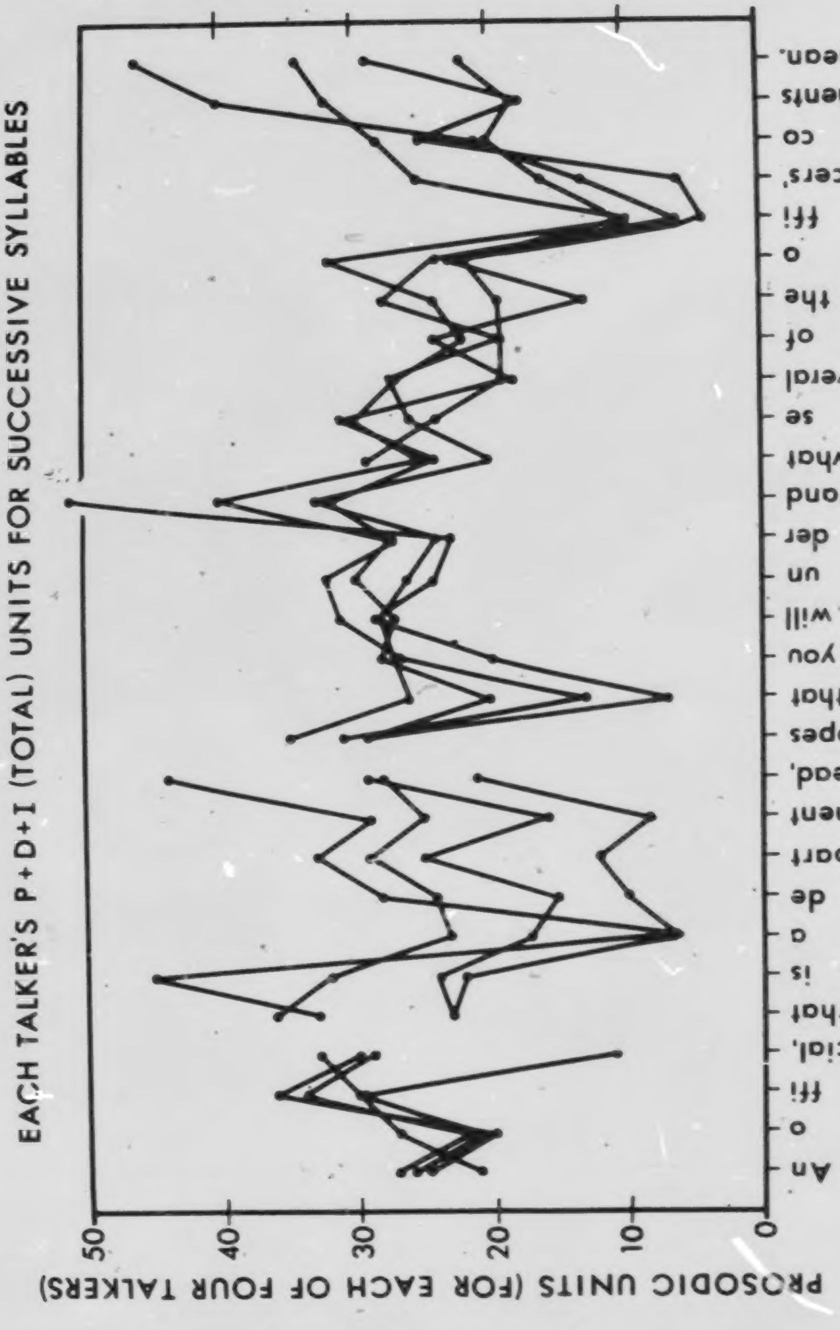

EACH TALKER'S P+D+I (TOTAL) UNITS FOR SUCCESSIVE SYLLABLES
PROSODIC UNITS (FOR EACH OF FOUR TALKERS)
50
40
30
20
10
0
An
o
ffi
cial,
that
is
a
de
part
ment
head,
hopes
that
you
will
un
der
stand
what
se
veral
of
the
o
ffi
cers'
co
mments
mean.

FIGURE 4

A further explanation presents itself if it is understood that lexical stress is based not only on the prosodic qualities of the syllables interior to a word, but also on the prosodic relationships between the word and its context, especially the relationships with adjacent syllables. Thus, if syllable A, e.g., [(k)ɑ], of a two syllable word A + B, e.g., [(k)ɑmɛn(s)], is preceded by a weak syllable in the previous word, then the stress of syllable A will be enhanced--and if syllable A is preceded by a weak syllable *and* followed by a comparatively small rise in syllable B, then the large positive rise leading up to A will probably override the small negative stress effect on A provided by the shallow rise from A to B, and A will be heard as more stressed than B. And, if rising syllable B, in the circumstance just described, is followed by a substantial rise in the next word (as is the case in [min] following the aberrant rising slopes seen in [mɛn]), the stress value of B will be further diminished. (These apparent contextual effects seem entirely logical, and a pilot perceptual test points to their validity, but they remain to be tested rigorously.)

To summarize Figure 4, we have seen that the acoustic contours (produced by summing the parameters in weighted prosodic units) reflect lexical stress patterns in nearly all the cases; only 4 of the 64 syllables (or 6 percent of the slopes) in the words of two or more syllables were "wrong." It may be inferred then, that larger prosodic patterns--such as phrasal stress patterns--are also as satisfactorily represented in the acoustic contours so derived. In short, although the measurement technique that has been described is cumbersome by the manual methods that were employed, it has utility for stress retrieval on the acoustic plane, and it should be reasonably simple to automate, *given preestablished syllable boundaries*.

Although the essence of our approach has been given in the preceding figures, and particularly in Figure 4, an additional view is presented in Figure 5 (top contour, in heavy line) in order to show the average of the four talkers' contours (from Figure 4), which can be considered the basic stress profile for the sentence. In the heavy contour all of the lexical patterns are correct in general shape, with the exception of syllable two in "comments," already discussed. Note, for example, the correct contrasts in the pattern of the word "official" versus "officers," and compare the contour of "understand" (stress pattern: mid, low, high) with that of "officers" (high, low, mid).

A few further observations can be made on the basis of this generalized contour (which is also representative of the contours found in the two longer sentences that were examined). (1) Local peaks in the contour are the relatively stressed syllables, and local valleys are unstressed syllables. (2) Peaks with the deepest valleys (immediately adjacent) are the stressed syllables of words that are high in information in the utterance. (3) The stressed syllable at a peak is usually part of a content word. (4) Relatively flat valleys consisting of two (or more) syllables are likely to contain at least one function word. (5) Rising slopes of the contours appear to contain more significant semantic information than falling slopes. (6) When the local peaks are connected by lines to produce a supralexical prosodic contour, the peaks then produced are the stressed syllables of words of major semantic importance in the utterance (i.e., key words).

It should be mentioned that a variety of phoneticians who have examined these data report that the contours shown in Figures 4 and 5 (top) closely resemble their intuitive impressions of the expected *intonation* pattern for the

AVERAGED TOTAL UNITS FOR FOUR TALKERS
and (below that)
AVERAGED PITCH vs DURATION vs INTENSITY FOR FOUR TALKERS

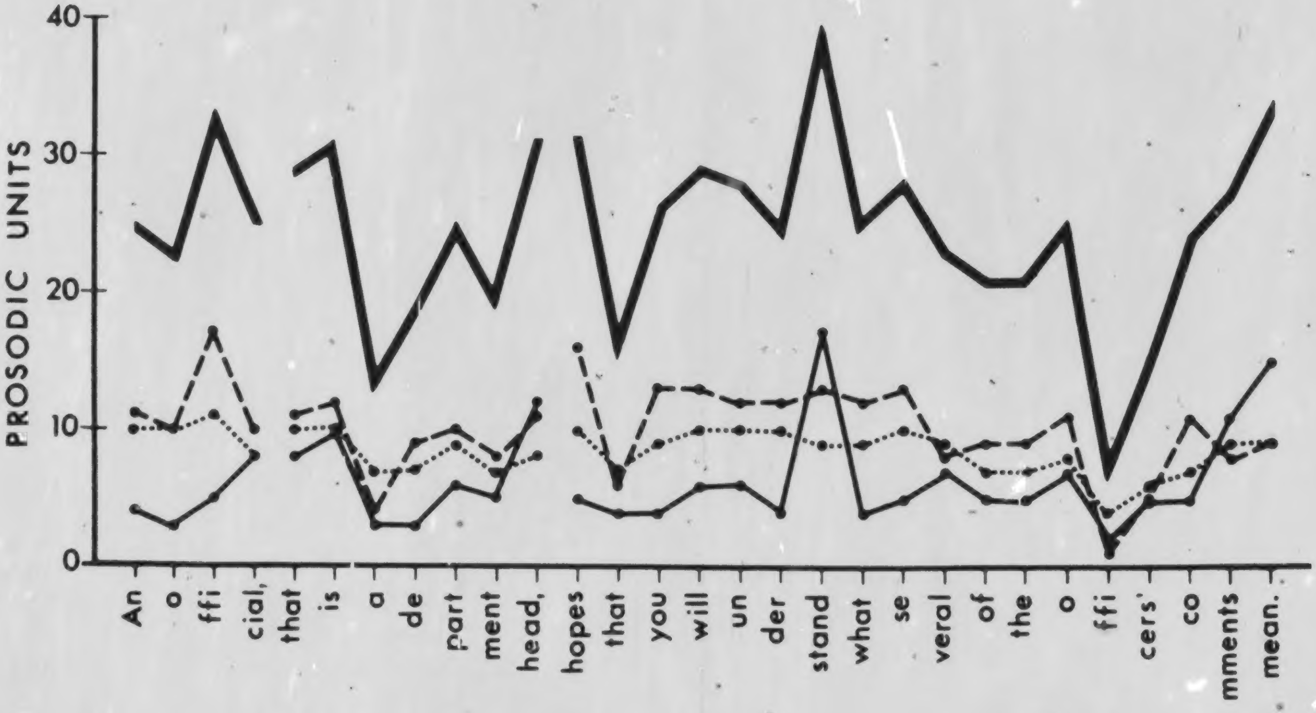


FIGURE 5

sentence. As has been demonstrated, "total syllable contours" are produced by summing the prosodic contents of the successive syllables, and therefore represent duration and intensity prominence effects in addition to those of pitch, although pitch (it will be recalled) is more heavily weighted in the display.

The averaged separate parameters are shown in the lower portion of Figure 5. It can be seen that all three run nearly parallel courses, except that duration rises prepausally, which is its normal configuration. Because these contours represent averaged data, it is not surprising to observe highly cooperative tendencies in slope across the parameters, but such profound agreement in the prosodic trio throughout an utterance is apparently less common in the speech of individual talkers (as has been illustrated in Figure 2).

CONCLUSION

The merit in using the procedure described is that it quite dependably elicits reasonable lexical stress patterns from limited acoustic information on connected speech.

We have also used a version of this technique in estimating (by eye) stress relationships in the syllables of unknown utterances--in spectrograms--with helpful results. However, the technique has an important prerequisite: the extent of syllable voicing must be known (i.e., syllable boundaries must be pre-established) in order to proceed with prosodic measurement, summing of syllable values, and the construction of contours. Several tests of this stress retrieval method have also been initiated using an algorithm for automatic syllable segmentation (Mermelstein and Kuhn, 1974) as a point of departure, and the results are promising. (It must be noted that segmentation of lexical boundaries, as such, is not attempted.)

It was observed above that a very few syllables (for individuals) failed to produce the expected prosodic contours. Two of these aberrant cases involved an unstressed syllable--containing a syllabic "ℓ"--that was long in voicing, following a stressed syllable that contained a short vowel, e.g., "official" [ə-Í-ℓ̩], "several" [É-vrℓ̩]. We are mindful of the fact that not only the voiced portions of speech, but also the voiceless regions, contribute to stress effects, even though, for the purpose of this experiment, the voiced portions were assumed to be the significant prosodic domain. In perceptual tests of stress that are being run, pairs of sequential syllables from the described sentence, spoken by individual talkers, are presented in two stimulus types in order to compare the prosodic contributions of the voiced portions alone (in one test) with those of whole syllables (in another test), and to compare the results of both of these test varieties with the acoustic contours derived as shown here. A pilot test, employing only the voiced portions of syllable pairs, indicates strongly that the acoustic contours--derived as in Figure 4--do, in fact, reflect perceived lexical and phrasal stress patterns. (Over a hundred syllable pairs from utterances by three of the four talkers have thus far been presented to six listeners.)

In the Introduction, it was mentioned that a total of three very long sentences were examined acoustically, of which the one that has been described above was the shortest. It is worth noting that the polysyllables that appear only once in this shortest sentence appeared at least twice in the three-sentence sample, and very similar lexical patterns--different only in degree--were elicited in the acoustic contours for each version of a given word, despite changed grammatical usage and/or sentence location.

REFERENCES

Bolinger, D. L. (1958) A theory of pitch accent in English. Word 14, 109-149. [Also in Bolinger, D. L. (1965) Forms of English: Accent, Morpheme, Order, ed. by I. Abe and T. Kanekiyo. (Cambridge, Mass.: Harvard University Press), pp. 17-55.]

Carhart, P. W. and J. S. Kenyon. (1934) A guide to pronunciation. In the Preface to Webster's New International Dictionary, 2nd ed. (New York: G. &. C. Merriam Co.), pp. xxxiv-xxxvii.

Coker, C. H., N. Umeda, and C. P. Bowman. (1973) Automatic synthesis from ordinary English text. IEEE Trans. Audio Electroacoust. AU-21, 293-298.

Denes, P. (1959) A preliminary investigation of certain aspects of intonation. Lang. Speech 2, 106-122.

Denes, P. and J. Milton-Williams. (1962) Further studies in intonation. Lang. Speech 5, 1-14.

Dewey, G. (1923) Relative Frequency of English Speech Sounds, 1950 rev. ed. (Cambridge, Mass.: Harvard University Press).

Fry, D. B. (1958) Experiments in the perception of stress. Lang. Speech 1, 126-152.

Gaitenby, J. H. (1965) The elastic word. Haskins Laboratories Status Report on Speech Research SR-2, 3.1-3.12. [Also published with four other papers from Haskins Laboratories Status Reports as Issledovanie Rechi (Speech Research). Translated into Russian by N. G. Zagorujko, USSR Academy of Sciences, Siberian Division, Institute of Mathematics, Novosibirsk, 1967.]

Klatt, D. H. (1974) Prediction of vowel duration in sentences. Unpublished manuscript, pp. 8, 11, 13.

Lehiste, I. (1970) Suprasegmentals. (Cambridge, Mass., and London, England: MIT Press), pp. 6-53.

Lieberman, P. (1967) Intonation, Perception, and Language. (Cambridge, Mass.: MIT Press).

McClean, M. D. and W. R. Tiffany. (1973) The acoustic parameters of stress in relation to syllable position, speech loudness and rate. Lang. Speech. 16, 283-290.

Mermelstein, P. and G. M. Kuhn. (1974) Segmentation of speech into syllabic units. J. Acoust. Soc. Amer., Suppl. 55, S22 (Abstract J11).

Peterson, G. E. and I. Lehiste. (1960) Duration of syllable nuclei in English. J. Acoust. Soc. Amer. 32, 693-703.

Thorndike, E. L. and I. Lorge. (1944) The Teacher's Word Book of 30,000 Words. (New York: Teachers College Press).

Is it VOT or a First-Formant Transition Detector?*

Leigh Lisker[+]
Haskins Laboratories, New Haven, Conn.

ABSTRACT

Discussion of voicing as a distinctive property of English stop consonants in initial position has centered on the measure of "VOT," the time of onset of laryngeal signal relative to the noise pulse generated by the stop release, but it has been shown that listeners' selection of b,d,g vs. p,t,k responses to synthetic stop + vowel stimuli is not determined entirely by VOT. Significant effects have been reported to depend on the behavior of the first-formant (F1) frequency immediately following voice onset, and on this basis it has been suggested that a feature detector responsive to a rapidly shifting F1 better explains the infant's discrimination of the two stop categories than some mechanism that measures VOT directly. The relative importance of VOT as against the presence vs. absence of F1 frequency shift after voice onset is assayed in several synthesis experiments in which VOT and F1 configurations are systematically varied. Labeling data obtained indicate that varying VOT regularly affects a significant change in listeners' judgments, and that varying F1 has some effect too; however, this latter variation is neither necessary nor sufficient to shift judgments decisively from one stop category to the other. The data further suggest that the presence of an F1 rising transition after voice onset serves as a voiced-stop cue not because of its dynamic aspect but simply because its onset frequency is low, i.e., at a value appropriate to a closed or almost closed state of the oral cavity.

The large and still growing literature on the phonetic features that serve to distinguish linguistically distinct categories of homorganic stop consonants has been very recently augmented by a short but interesting contribution from Stevens and Klatt (1974). The burden of their report is that perceptual importance attaches to the fact that for the voiceless aspirated stops of English the onset of voicing associated with a following stressed vowel occurs at about the time that the first formant has achieved the frequency appropriate to that

---

*Paper presented at the annual meeting of the American Association of Phonetic Sciences, St. Louis, Mo., 5 November 1974.

[+] Also University of Pennsylvania, Philadelphia.

[HASKINS LABORATORIES: Status Report on Speech Research SR-41 (1975)]

vowel. Thus the so-called voice onset time (VOT) measure, i.e., the duration of the interval between onset of the burst resulting from stop release and onset of glottal signal, has a value essentially equal to the duration of the oral opening gesture. By contrast, English /b,d,g/ are characterized by VOT values such that the formant transitions following release are excited by the glottal source over a significant portion of their total duration. On the basis of certain data from experiments in synthesis, it can be demonstrated that the boundary along the VOT dimension between /d/ and /t/ is not completely stable but may vary considerably as a function of the rate and/or the duration of the transition. Of five subjects tested, one appeared to be responding more as though measuring the interval from release to voice onset, while another's responses were to the interval between voice onset and the specific time at which the formant transition was completed. The other subjects were intermediate between these two, i.e., they seemed to use a mixture of these two strategies. On the basis of this finding, Stevens and Klatt (1974) suggest that listeners generally have the ability to respond differentially to signals depending on whether or not they present a pulse-excited first formant of rapidly shifting frequency. Furthermore, they suggest that this ability, rather than one that "simply" measures VOT, is what the language-acquiring infant relies on in the first steps toward a mastery of English phonology.[1] The measure proposed by Stevens and Klatt is a kind of complement to VOT, namely the transition duration minus VOT, and we might accordingly call it simply "VTD" for "voiced transition duration." Voiced transition duration has the merit that it very probably is more independent of place of stop articulation than is VOT, since it appears that VOT and burst and transition durations all increase from labial to alveolar to velar place of closure. This would seem to say that, inasmuch as speech production is for perception, the English talker controls the timing of voice onset not with reference to the stop release, but rather in relation to the achievement of the steady-state vowel target formant frequencies. Of course if there is no very significant variation in transition, at least for a given place of stop articulation before a given vowel, one might even suppose that the talker times the onset of voicing in relation to release, but that the listener attends to whether or not there is movement of the first formant after voicing onset.

A reading of the literature concerned with the acoustic cues to stop voicing indicates that there should be nothing surprising about the finding that the F1 transition plays a role: a very early paper on speech synthesis (Cooper, Delattre, Liberman, Boret, and Gerstman, 1952) reported that "the transitions of the first formant appear to contribute to voicing of the stop consonants" (p. 600). Nor should it be thought at all extraordinary to find still other acoustic features--fundamental frequency contour, for example--that also control to some extent the phonetic classification of stop patterns as voiced or voiceless. What would, in fact, be much more difficult to justify would be an assertion that any particular feature isolable in the acoustic signal plays absolutely no role in the listener's phonetic categorizations. Certainly, some such features play a vanishingly small role, but given the experimental strategies used in discovering the acoustic cues, it is hard to imagine a feature not utterly imperceptible that could be shown to have absolutely zero cue value. The question that can reasonably be asked is: What is the relative importance of one feature compared with others? If it is claimed, for example, as Haggard, Ambler, and Callow (1970) apparently do,[2] that fundamental frequency has an importance of the order that may be claimed for VOT, one might ask whether the two features are equally necessary or perhaps equally sufficient as cues to the contrast, or whether there are F0 contours for which varying VOT has no effect on labeling behavior, even as there appear to be values of VOT for which varying the F0 contour has no effect on voicing judgments. The same questions can be raised with respect to the Stevens-Klatt hypothesis if, as is reasonably inferred from their argument, they mean to claim for the VTD feature a perceptual importance equal to that determined for VOT.

---

[1] Eimas, Siqueland, Jusczyk, and Vigorito (1971) have pointed out that that same infant can, like the adult English speaker, distinguish a VOT of +20 msec from one of +40 msec.

In the earliest work in this area done at the Haskins Laboratories, the pattern feature isolated for primary attention was called "first-formant cutback"; in later studies the preferred term was "VOT." In all these studies the point was made, more or less insistently, that first-formant attenuation before voicing onset and the timing of that onset were to be thought of as acoustic features that together were manifestations of a shift in laryngeal state from a wide-open and nonvibrating to a closed-down and vibrating glottis. The terminological shift from "F1 cutback" to "VOT" was occasioned by a shift of attention from the perceptual evaluation of synthetic speech patterns to the precise measurement of spectrographic patterns of human vocal-tract speech and to the underlying physiological and articulatory events. In spectrograms of natural speech, F1 cutback is simply very hard to measure; it is not easy to determine the exact time at which F1 reaches full amplitude nor do spectrograms suggest that the F1 amplitude is all that stable. The VOT measure, although it has its difficulties, to be sure, is much more easily accomplished, and by now the published data leave little ground for doubting its usefulness as a basis for distinguishing between stop categories. I would guess that the Stevens-Klatt measure of VTD, which would require fixing both the time when F1 reaches some criterial intensity level and when it reaches the steady-state frequency of the following vowel, is not one that will be attempted for any large number of spectrograms of natural speech. F1 cutback and VTD are easily measured for synthetic speech patterns when those patterns are fabricated with these measures in mind. If the human listener had only to contend with such patterns, it would be so much simpler to describe speech perception. In the case of F1 cutback and VTD the match between natural and synthetic speech patterns is not easily accomplished, for the reasons just stated; in the case of VOT a very close match indeed has been determined, both for English and several other languages as well.[3] I think the question of determining the match between natural speech and synthetic is an important one, for we know that the match need not be slavishly close for a synthetic stimulus set to be a perceptually satisfactory match--at least from the gross phonemic labeling behavior aspect--to a set of natural speech utterance tokens. It may be remembered, for example, that a quite unnatural set of stimuli "accounted for" the /do/-/to/ contrast by varying F1 cutback alone, that is, with both VOT and VTD constant and, in fact, equal to zero (Liberman, Delattre, and Cooper, 1958). Evidently, this means that we may make inferences about the speech-handling capabilities of the sensory-perceptual system from the data of experiments in speech synthesis. At the same time, it indicates that we must be cautious in asserting just how these capabilities are exercised when natural speech signals are being processed.

---

[2] Haggard et al. (1970) report findings for which they provide no very clear interpretation. The fundamental frequency contour is said to serve as a stop category cue in synthetic speech patterns that are described as "ambiguous between /bi/ and /pi/" (p. 613), and, while not all subjects responded unequivocally to the stimulus set, the authors express the belief that F1 cutback is possibly no more robust a cue. They make no reference to VOT, and neither F1 cutback nor VOT values are specified for their test stimuli. A clearer picture of the relation between F0 contour and VOT is presented in Fujimura (1971). From his data Fujimura concludes that F0 plays a subsidiary role in the voiced-voiceless distinction among English initial stops.

Let us return to the question that serves as the title of our discussion. One possibly objectionable implication of that question is that one of the feature dimensions, VOT or VTD, plays little or no role in the voicing contrast, but it is quite reasonable to ask whether VOT and VTD is more important in some sense. It is this question Stevens and Klatt (1974) seem to have answered in favor of VTD, at least as a basis for understanding the behavior of Eimas's infant subjects. I think there are grounds, in particular the data represented in Figures 1 and 2 here, for believing that their VTD measure has less significance than they would assign to it.

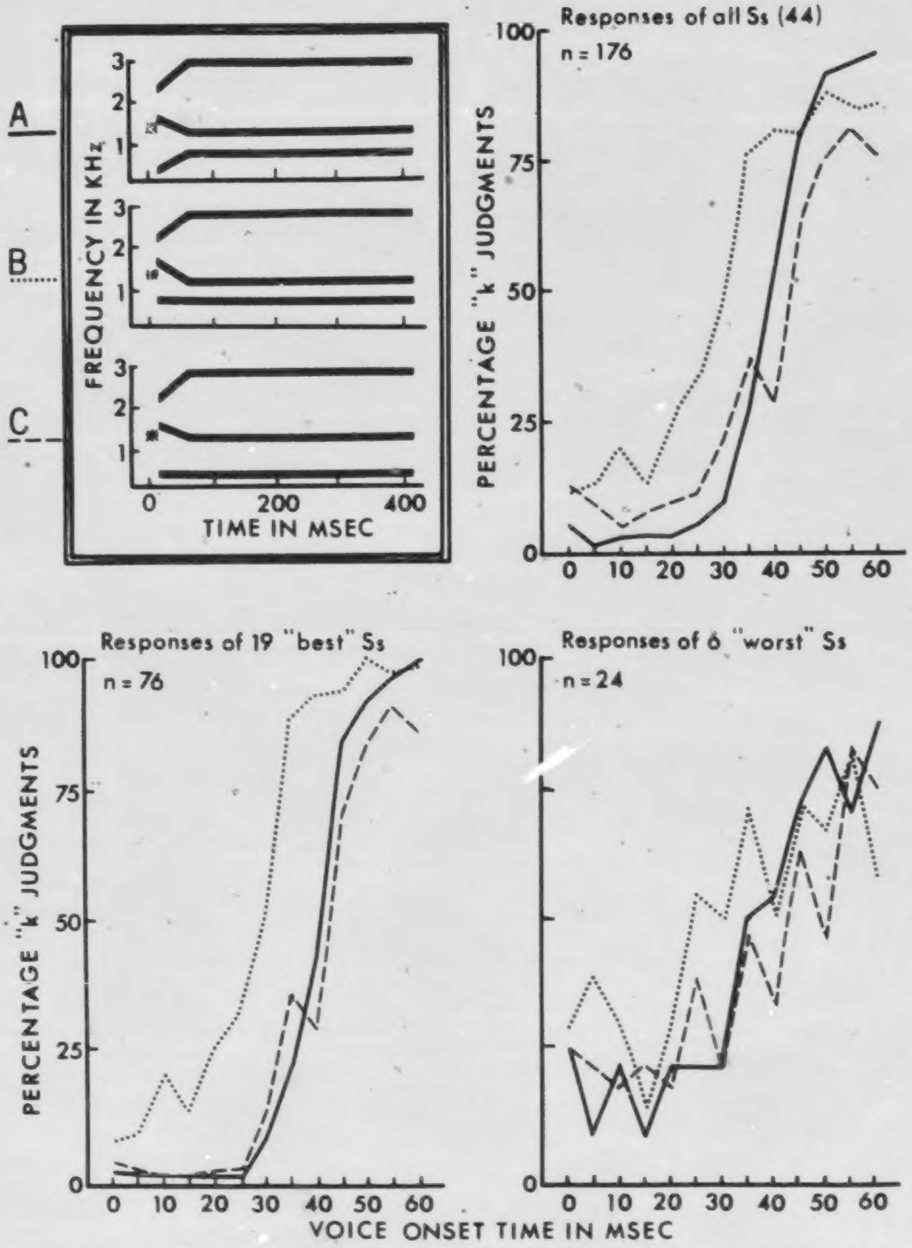
Figure 1 represents the labeling responses of 44 phonetically naive University of Connecticut students to the type of stimuli shown schematically in the upper left-hand quadrant. Stimulus type A is composed of a burst and formant-transition configuration appropriate to the velar stop place of articulation, and the transition is followed by a steady-state formant pattern heard as the vowel /a/. From this basic pattern a set of 13 stimuli was generated (with the help of the Haskins Laboratories parallel resonance synthesizer under computer control) by varying VOT together with F1 onset from a value of 0 to +60 msec in steps of 5 msec. Burst and transition durations were fixed at 20 and 45 msec, respectively. The solid curve in the upper right-hand quadrant of the figure represents percentage /k/ responses as a function of VOT for all 44 subjects tested. The test was the usual "forced choice" one, with responses restricted to /g/ and /k/. The point at which responses were divided evenly between /g/ and /k/ falls at just about VOT = +40 msec.

In the lower left- and right-hand quadrants of Figure 1 are shown the responses of the 19 "best" subjects, those who labeled the largest number of stimuli identically on 4 exposures, and the 6 "worst" subjects, who were most nearly random in behavior. Even the worst subjects show a crossover value between /g/ and /k/ along the VOT dimension, at about +35 msec.

All the responses to the type A stimuli can be compared, first of all, with those elicited by patterns of type B, which differ from A only in that the first

---

[3] Data for English, French, Spanish, Thai, and Korean speakers can be found in one or more of the following: Abramson and Lisker (1965, 1972, 1973); Lisker and Abramson (1970); Caramazza, Yeni-Komshian, Zurif, and Carbone (1973); Williams (1974).



The k-g Contrast: VOT vs First-formant Frequency

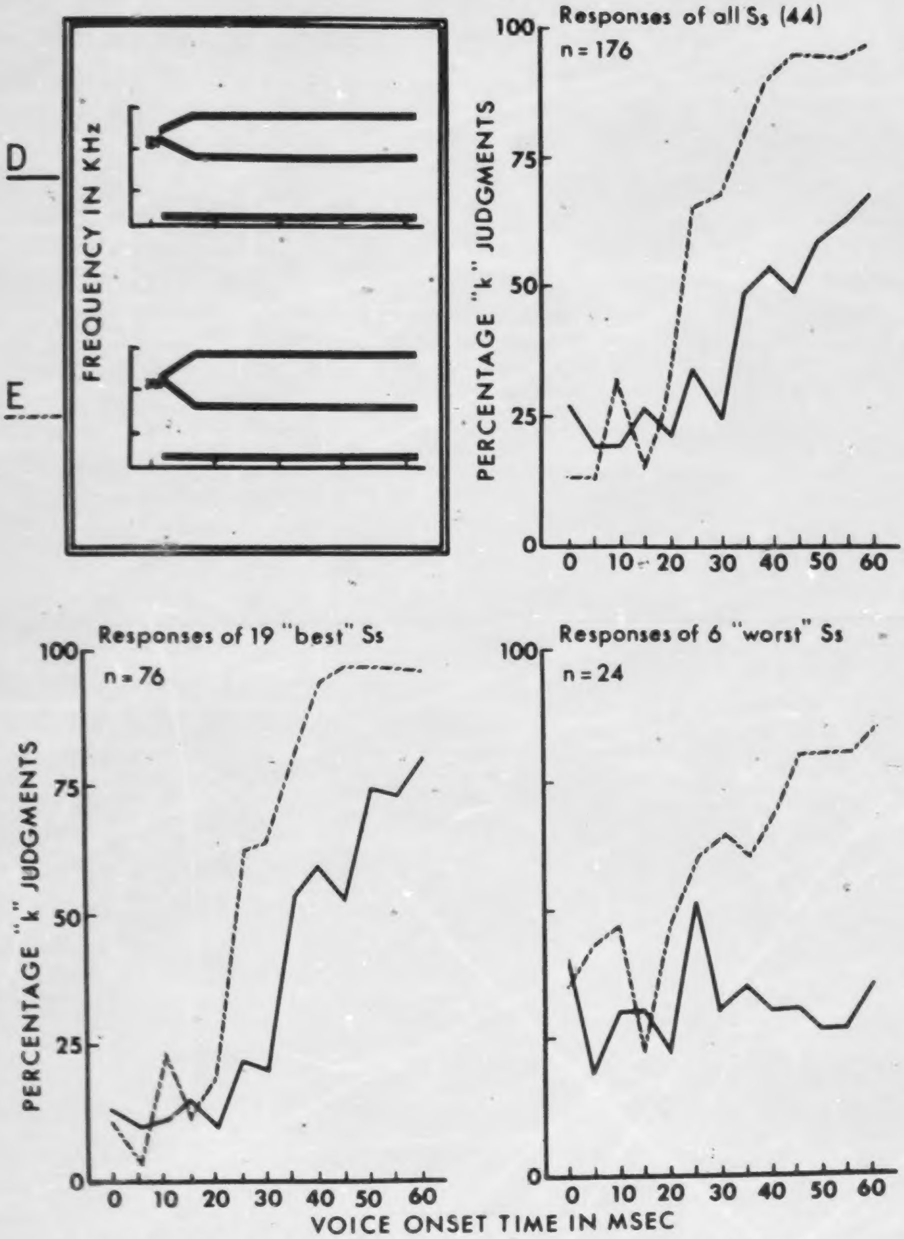
(n represents number of judgments for each data point)

FIGURE 1

formant has its frequency fixed at the steady-state value of the /a/, i.e., 769 Hz in these particular patterns. The dotted line showing the responses to the B stimuli indicates that the presence of a sharply rising F1 is not a requirement for a majority of subjects to report hearing /g/; even the six "worst" subjects gave mostly /g/ responses for VOT less than +25 msec. Certainly, for the 19 "best" subjects, it appears that responses to B stimuli consistently show more /k/ judgments over the entire VOT range than for the A patterns; the B curve is displaced to the left of the A curve by about 10 msec. Moreover, while /k/ judgments reach 100 percent for large VOT values, /g/ judgments are no better than 90 percent at VOT = 0. If we ask whether there is any VOT value for which the pattern difference between A and B is sufficient to shift judgments from mostly /g/ to mostly /k/, the answer is that, for all 44 subjects, there is precisely one value of VOT, namely +35 msec, at which pattern A elicited mainly /g/ responses (73 percent) and pattern B mostly /k/ responses (76 percent). For all other values of VOT the two patterns were judged, by a greater or lesser majority, to belong to the same stop category. For the 19 "best" subjects the A patterns with VOT = +35 msec yielded 79 percent /g/, while the B pattern with the same VOT value was scored 88 percent /k/.

Pattern C resembles B in having a straight first formant; it differs in that the frequency of that formant is very near (386 Hz) the onset frequency of the bent F1 of pattern A (361 Hz). The effect of this lowering of the F1 onset frequency is seen most dramatically in the responses of the "best" subjects: for small VOT values as many /g/ responses were elicited by pattern C as by A, despite the absence of any F1 frequency shift in C. In fact, it would seem as though pattern C differs from A mainly in that at the higher VOT values it yielded somewhat fewer /k/ judgments. In other words, it might be said that the lower steady-state F1 frequency is a more strongly pro-/g/ cue than the absence of an F1 frequency shift is pro-/k/. It must, of course, be conceded that pattern C, with post-transition F1 and F2 frequencies of 386 and 1282 Hz, respectively, is heard as a stop followed by a vowel other than /a/, but we must presume that a theory of stop voicing perception must, to be adequate, be able to account for more than a single vowel context. The data for patterns A, B, and C suggest that it is not so much F1 frequency shift as simply F1 onset frequency that favors /g/. A low F1 frequency tells the listener that the mouth is not very open, whether or not it is very soon to be more open.

Figure 2 presents labeling data for patterns whose post-transition first and second formants have frequencies other than those of the previous patterns. Pattern D, with a straight F1 at 286 Hz, yielded a lower percentage of /k/ judgments than any of the other patterns tested; the six "worst" subjects, in fact, gave mostly /g/ responses for all but a single value of VOT. This behavior is understandable if we suppose that the low onset frequency of F1 is a strong voicing cue. However, the failure of the "worst" subjects to report /k/ for high VOT values is troublesome. A possible explanation is that the vowel quality was bizarre to the point where there was a complete failure to identify the consonant-vowel (CV) sequence, and consequently these subjects were unable to pick up any of the features they attended to in the other patterns and were simply giving random responses. Of course, pattern D differs from those previously discussed in having an F2 whose steady-state frequency is considerably higher, and we might entertain the notion, harebrained on the face of it, that the raised F2 is the cause of this massive shift to /g/ judgments. Pattern E disposes of such a hypothesis, however, since its second formant is almost as high in frequency. With its steady-state F1 at the midrange value of 413 Hz, pattern E yielded very



The k-g Contrast: VOT vs First-formant Frequency

(n represents number of judgments for each data point)

FIGURE 2

solidly /k/ responses for high VOT values, while at the low end of the VOT range there was a preponderance of /g/ responses, though no more than for D and somewhat less than for A, B, or C. In terms of crossover values along VOT there are differences among the five patterns tested: for all subjects the crossover is earliest for pattern E and latest for C, the difference being 20 msec. The most obvious difference between these patterns is in F2 frequency, but to consider this the basis for the response difference means to suppose that raising F2 increases /k/ judgments, and this makes no more sense, on the face of it, than the contrary hypothesis generated by the comparison of responses to patterns C and D. The only thing left to say is that I have nothing plausible to suggest in explanation, and that work is continuing.

Figure 3 shows labeling data in response to sets of stimuli all having transitions in which the first, second, and third formants are rising, so that responses were either /ba/ or /pa/. The variable was transition duration, and the purpose was to replicate the Stevens-Klatt experiment with /da/-/ta/ patterns. The results are very similar to those reported in that study; there is a shift in VOT crossover of just about 30 msec for a 50-msec change in transition duration. One additional observation is perhaps worth making. As the duration of transition decreases, the VOT crossover value decreases by a slightly lesser amount, but the change from one of 25 msec to the shortest duration of transition tested has no effect on the crossover value, which remains at slightly greater than +20 msec, a familiar value for the labial place of articulation.

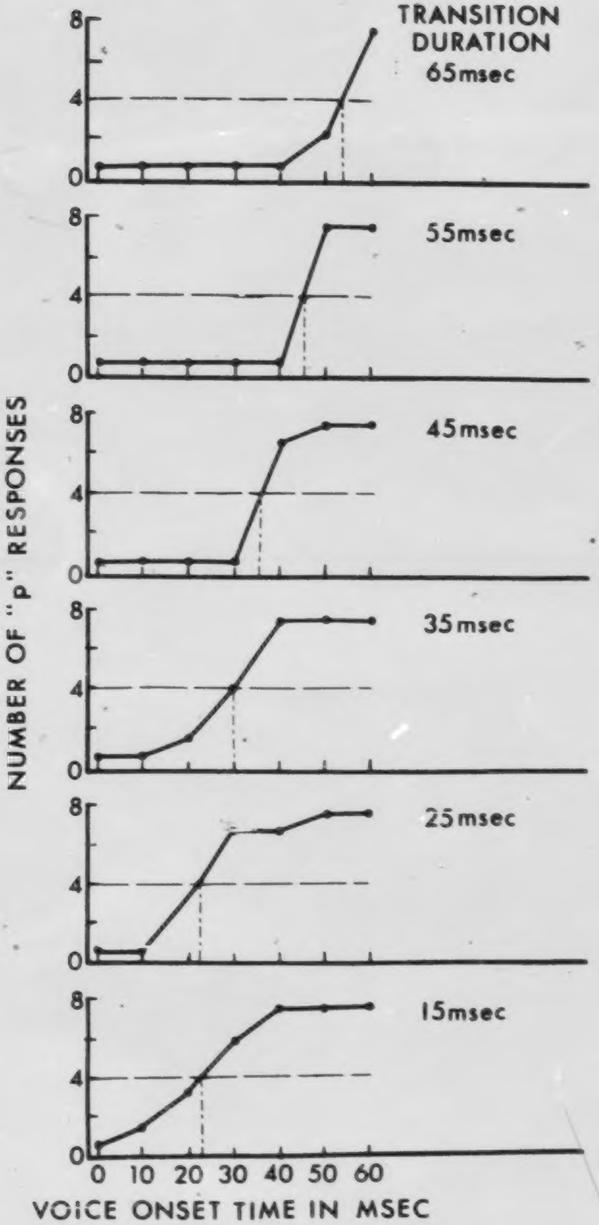
The data displayed in Figure 4 are meant to answer this question: Does the extent of F1 frequency shift effect any significant shift in VOT crossover value? The patterns tested had transitions like those of the previous set, but their transition durations were fixed at 45 msec, and F1 had a fixed onset frequency of 154 Hz and then rose linearly to a steady-state frequency whose value was varied over the stimulus set from 260 to 769 Hz, in steps of roughly 100 Hz.[4] Needless to say, with F2 fixed at a post-transition value of 1620 Hz, the patterns were judged by our American subject to contain vowels of a most peculiar kind. The display suggests that while the crossover value wavers somewhat over a range smaller than 10 msec, there is no systematic shift with increasing extent of F1 transition.

The data represented in Figures 3 and 4, unlike those shown in the first two figures, were derived from only a single subject, and this fact may explain certain discrepancies among the different data sets that a closer examination than is warranted here would bring out.

To sum up, our data suggest that the presence of a voiced F1 transition is not a requirement for stops to be heard as /b,d,g/. None of our experiments discussed here tell us, to be sure, whether absence of voiced F1 transition is a requirement for English initial /p,t,k/. Of course, a pattern with VOT equal, let us say, to +50 msec and with F1 beginning at that point with a low frequency and rising transition would hardly be found in natural speech. More to the point, however, is the fact that from experimental data not yet quite ready for presentation it appears that such patterns are not heard as /b,d,g/ + vowel, but p versus b

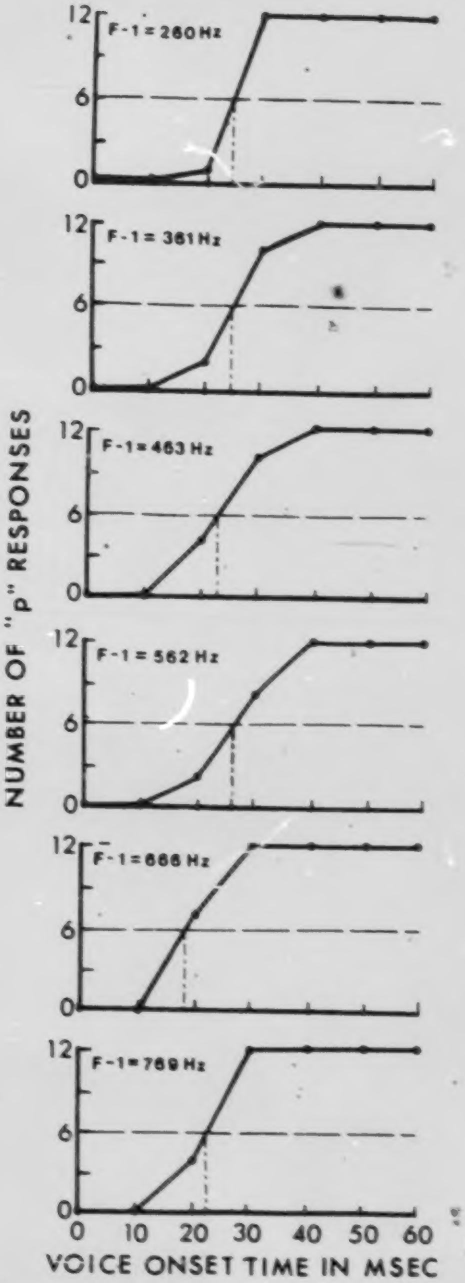
---

[4] The other way of varying F1 transition extent, namely, by varying the onset frequency, is known to affect stop voicing judgments [see Cooper et al. (1952)].



Voice Onset Time versus Transition Duration

FIGURE 3



The p-b Contrast: VOT vs F-1 "Target" Frequency

FIGURE 4

as /p,t,k/ + some other phonetic segment (perhaps /l/ more than anything else) + vowel. A sharply rising F1, moreover, is most likely to be found in sequences of stop and a vowel with a high F1; with the vowels /i/ and /u/ such a feature is much less evident. Unless infants learn their stop voicing distinctions primarily from exposure to stops before the vowels /a/ or /æ/ (and perhaps they do!), it seems doubtful that VTD--certainly a highly context-sensitive dimension--triggers a built-in device, while the much less context-sensitive VOT[5] does not. Moreover, the notion that VTD triggers a basic mechanism, for all its appeal, suggests that English owes its important position in the present-day world to the fact that very many languages seem perversely *not* to exploit it: Spanish-speaking children, for example, must presumably learn to ignore information provided by this F1 transition detector as one aspect of their process of language acquisition. If we say that the English-speaking learner calls the initial prestress stop /p,t,k/ if he detects aspiration, and that his detection of this feature rests significantly on the absence of F1 transition after voice onset, then we may ask why Hindi-speaking listeners seem to require longer VOT than do American listeners before they report hearing voiceless aspirated stops.[6] It is not necessary to look far afield for languages that do not exploit the VTD dimension; English itself contrasts voiceless inaspirates and voiced stops medially, and VOT does a fair job of separating them. Where VOT fails, VTD does not help.

REFERENCES

Abramson, A. S. and L. Lisker. (1965) Voice onset time in stop consonants: Acoustic analysis and synthesis. In *Proceedings of the 5th International Congress on Acoustics*. (Liège: Imp. G. Thone), *A51*.

Abramson, A. S. and L. Lisker. (1972) Voice timing in Korean stops. In *Proceedings of the 7th International Congress of Phonetic Sciences*, Montreal, August 1971. (The Hague: Mouton), pp. 439-446.

Abramson, A. S. and L. Lisker. (1973) Voice-timing perception in Spanish word-initial stops. J. Phonetics *1*, 1-8.

Caramazza, A., G. H. Yeni-Komshian, E. B. Zurif, and E. Carbone. (1973) The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals. J. Acoust. Soc. Amer. *54*, 421-428.

Cooper, F. S., P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman. (1952) Some experiments on the perception of synthetic speech sounds. J. Acoust. Soc. Amer. *24*, 597-608.

Cooper, W. E. (1974) Contingent feature analysis in speech perception. Percept. Psychophys. *16*, 201-204.

---

[5] This is controversial, however. Lisker and Abramson (1967:15) claim that VOT is unaffected by the particular vowel following the stop, but a contrary finding is reported by Klatt (1973). In agreement with Klatt is Cooper (1974), who reports a somewhat higher VOT crossover value for /bi/-/pi/ than for /ba/-/pa/ in experiments with synthesized syllables. The Cooper data are difficult to interpret in the absence of detailed information about the transitional configurations involved.

[6] Statements to this effect by Hindi-speaking subjects are not in themselves strong evidence, but they are consistent with VOT measurements on Hindi reported in Lisker and Abramson (1964).

Eimas, P. D., E. R. Siqueland, P. Jusczyk, and J. Vigorito. (1971) Speech perception in infants. Science 171, 303-306.

Fujimura, O. (1971) Remarks on stop consonants: Synthesis experiments and acoustic cues. In *Form and Substance: Phonetic and Linguistic Papers Presented to Eli Fischer-Jørgensen*, ed. by L. L. Hammerich, R. Jakobson, and E. Zwirner. (Copenhagen: Academisk Forlag).

Haggard, M., S. Ambler, and M. Callow. (1970) Pitch as a voicing cue. J. Acoust. Soc. Amer. 47, 613-617.

Klatt, D. (1973) Voice-onset time, frication and aspiration in word-initial consonant clusters. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 109, 124-136.

Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1958) Some cues for the distinction between voiced and voiceless stops in initial position. Lang. Speech 1, 153-167.

Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word 20, 384-422.

Lisker, L. and A. S. Abramson. (1967) Some effects of context on voice onset time in English stops. Lang. Speech 10, 1-28.

Lisker, L. and A. S. Abramson. (1970) The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the 6th International Congress of Phonetic Sciences*. (Prague: Academia), pp. 563-567.

Stevens, K. N. and D. H. Klatt. (1974) Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Amer. 55, 653-659.

Williams, L. (1974) Speech perception and production as a function of exposure to a second language. Unpublished Ph.D. dissertation, Harvard University.

Pitch in the Perception of Voicing States in Thai: Diachronic Implications*

Arthur S. Abramson[+]
Haskins Laboratories, New Haven, Conn.

It is tempting for the experimental phonetician to believe that phonetic hypotheses on the causes of sound change should be testable in the laboratory. In the absence of any technological innovation that allows the resurrection of long-dead informants for ever so brief a stint of field work, perhaps the most we can hope to do is to test the phonetic plausibility of these hypotheses by using present-day speakers of one or more of the languages concerned. Even if little light is shed on the historical process, new information on the phonetic nature of the phonological categories of interest may be added to the literature. This paper represents just such an attempt. It examines changes in stop consonant voicing in the Tai family of languages by seeking new information on acoustic cues in modern Thai.

Changes in stop voicing must be viewed against the background of the putative emergence of tones in the Tai family (Gedney, 1974) as a function of initial consonants. Many scholars, e.g., Maspéro (1911), Li (1947), and Coedès (1949), have argued that for Tai and other families of Southeast Asia low tones have developed in word classes with ancient voiced initials, and high tones have developed in word classes with voiceless initials. Such an argument has at least indirect support from acoustic phonetic research, principally on English. House and Fairbanks (1953), as well as Lehiste and Peterson (1961), showed that the fundamental frequency (f_0) of phonation soon after the release of an English voiceless consonant is higher than after a voiced consonant. The phonetic rationale for the effect is that the unimpeded air flowing through the open glottis for the voiceless consonant momentarily perturbs the vibration rate of the vocal folds upward, once voicing begins, while the somewhat impeded air flow of the essentially closed glottis for voiced consonants may provide insufficient force to keep the vibration rate at the intended level and thus allow a slight drop in frequency during and just after the consonant closure or constriction. Recovery from the f_0 perturbation can take longer than the

---

*Presented at the annual meeting of the Linguistic Society of America, New York, 27-30 December 1974.

[+] Also University of Connecticut, Storrs.

Acknowledgment: This research was conducted while the author was on sabbatical leave in Thailand on research fellowships from the American Council of Learned Societies and the Ford Foundation Southeast Asia Fellowship Program. I gratefully acknowledge the hospitality of the Faculty of Humanities, Ramkhamhaeng University, and the Central Institute of English Language, both in Bangkok.

[HASKINS LABORATORIES: Status Report on Speech Research SR-41 (1975)]

probable duration of the aerodynamic perturbing factor itself. We may suppose that one transient is caused by the disturbing force, and a second transient is manifested by the f_o movement from its momentary excursion back toward the intended contour of the syllable.

Modern Thai (Siamese) has three categories of stop consonants, usually called voiced, voiceless unaspirated, and voiceless aspirated. In this study I have taken the labial stops to represent the system and concentrated on them. Recently Erickson (in press) has found that the voiced labial stop of Thai typically shows a low f_o, while the two voiceless stops show high values. As for the latter pair, the aspirated stop tends to have a higher value than the unaspirated stop. [While agreeing with her general findings on the voiced-voiceless distinction, in recent work Gandour (1974) surprisingly finds that the aspirated stops show a smaller upward swing of f_o than do the unaspirated stops.] The notion is that in Proto-Tai such adjustments of f_o were heard as pitch perturbations that were gradually enhanced in speech until they achieved phonemic status. The argument would apply whether we are supposing a pristine state of tonelessness in Proto-Tai or indeed a phonology that already included, say, two tones, since the daughter languages today have five or more tones.

The Proto-Tai system of stop consonants, epitomized by the labials, is shown in Table 1. These reconstructions and their subsequent changes to the stop consonants of modern Thai represent the consensus of most scholars, except that there are serious questions about the phonetic nature of the so-called glottalized b̲. Haudricourt and Martinet (1946), incidentally, posit murmured or voiced aspirated /*bh/ as an intermediate stage between /*b/ and /ph/.

TABLE 1

Proto-Tai:	*ˀb	*b	*p	*ph
Thai (Siamese):	b	ph	p	ph
Script:	บ	พ	ป	ผ

Using techniques of speech synthesis, other investigators--Fujimura (1971) for Japanese and English, and Haggard, Ambler, and Callow (1970) for English--have shown that pitch shifts can influence auditory judgments as to the voicing assignments of syllable-initial stop consonants. In addition, Lea (1973) has been very successful in using a f_o criterion in making voicing identifications of consonants in acoustic analysis of English utterances. To the best of my knowledge, such experiments, particularly perceptual ones, have not been tried for languages with more than two consonant categories distinguished by voicing features.

My plan was to see whether pitch shifts, brought about by control of the f_o parameter of a speech synthesizer, would affect listeners' judgments as to the voicing-class membership of initial stops. By way of background, it must be said that some years ago Lisker and I (Lisker and Abramson, 1964; Abramson and Lisker, 1965) showed, both acoustically and perceptually, that the three stop categories of Thai lie along the dimension of voice onset time, namely, the temporal relation between the closing of the glottis for audible pulsing and the release of the occlusion of the initial stop. To furnish a baseline for the present research, it was necessary to replicate the perceptual part of this old study with the new subjects who were to be used for the experiments on the efficacy of fundamental frequency perturbations. I used the Haskins Laboratories' parallel resonance synthesizer to produce a syllable of the type labial stop plus [a:]. Thirty-seven variants of the syllable were made to form a continuum of voice onset time ranging from a voicing lead of 150 msec, before the release of the stop, to a voicing lag of 150 msec, after the release. The range was divided into 10-msec steps except for the portion from a lead of 10 msec to a lag of 50 msec, which was divided into 5-msec steps. For voicing lead, I simply had low-frequency harmonics during the simulated stop occlusion. For voicing lag, during the interval after the release when no voicing is present, the second and third formants were filled with noise to simulate aspiration and the first formant was simply omitted to simulate the extreme consequence of an open glottis. I also tapered the overall amplitude in ways roughly appropriate to the effects of laryngeal timing. I restricted the experiment to the mid tone of the five tones of Thai by providing a flat fundamental frequency contour except for a slight dip at the end. The identification data for 48 native speakers of Thai are presented in Figure 1. These subjects were presented with the stimuli randomized into eight test orders for labeling as initial stop consonants. The ordinate shows percent identification. The abscissa shows values of voice onset time. Voicing lead is indicated in negative numbers, voicing lag in positive numbers, while zero means voice onset at the moment of release. The three expected categories emerge, although the middle one, unaspirated p, loses responses to the two categories on either side and does not get as close to 100 percent. The 50 percent crossover values between categories fall at -7 and +26 msec.

With the sufficiency of voice onset time as a cue once again demonstrated for Thai, I went on to new experiments. Unlike Haggard et al. (1970), who used f_o excursions far greater than any observed in the literature, I restricted my range to 20 Hz above a reference level and 20 Hz below. This choice is well in accord with Erickson's (in press) values for nine speakers of Thai. She found five male and four female adults to produce f_o perturbations for stops well within a range of 40 Hz and just one female with a range of 52 Hz. I set the level portion of my mid tone at 120 Hz and shifted upward to it from 110 and 100 Hz and downward to it from 130 and 140 Hz. For these f_o shifts I used three time spans: 50, 100, and 150 msec. I also made variants with no f_o shift, that is, a level f_o onset. Finally, for all these conditions I provided 13 voice onset time variants, the ones shown along the bottom of the graph in Figure 2. These values were chosen by pretests and inspection of the data of Figure 1. Thus for each voice onset time value there were 13 f_o variants, a flat one plus 12 perturbations, yielding 117 stimuli that were randomized eight times with a sample at the beginning of each tape.

The labeling responses of 46 subjects (two having dropped out) are given for the flat f_o onsets in Figure 2. This set is presented alone to make it easier to look at the rest of the graphs. The voice onset time values are arrayed along the bottom, while the percentages are at the left end. The bars are coded to show responses in terms of the three initial consonants. If we extrapolate from these bars, the results accord well with the perceptual crossover points in the preceding graph, falling at -10 msec and +22 msec. The responses to all of the f_o perturbations for each of the three time spans, 50, 100, and 150 msec, are given in Figures 3-5, respectively. From top to bottom on each page there are graphs for the four frequency shifts.

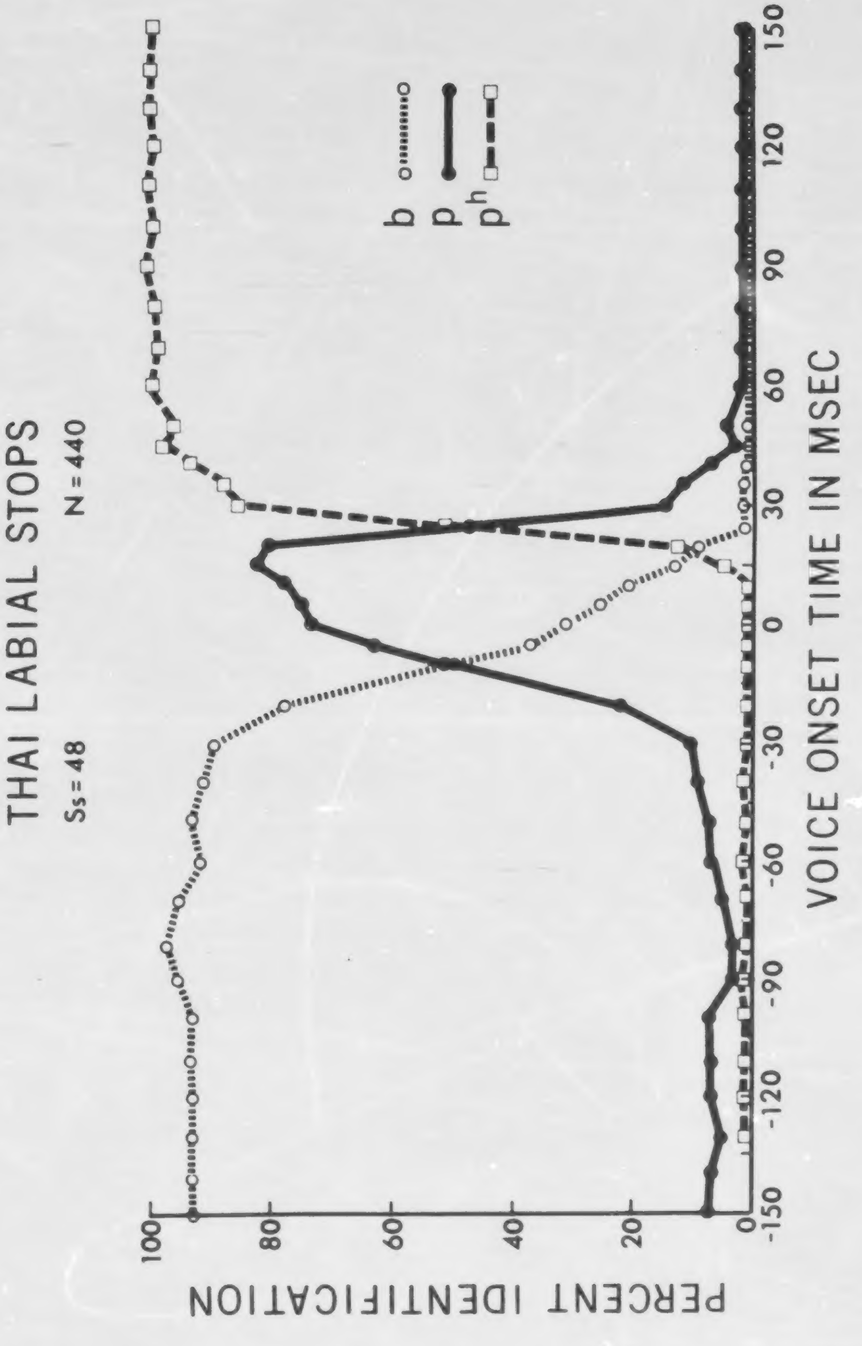

THAI LABIAL STOPS
Ss = 48
N = 440
PERCENT IDENTIFICATION
0
20
40
60
80
100
-150
-120
-90
-60
-30
0
30
60
90
120
150
VOICE ONSET TIME IN MSEC
b
p
p^h

FIGURE 1

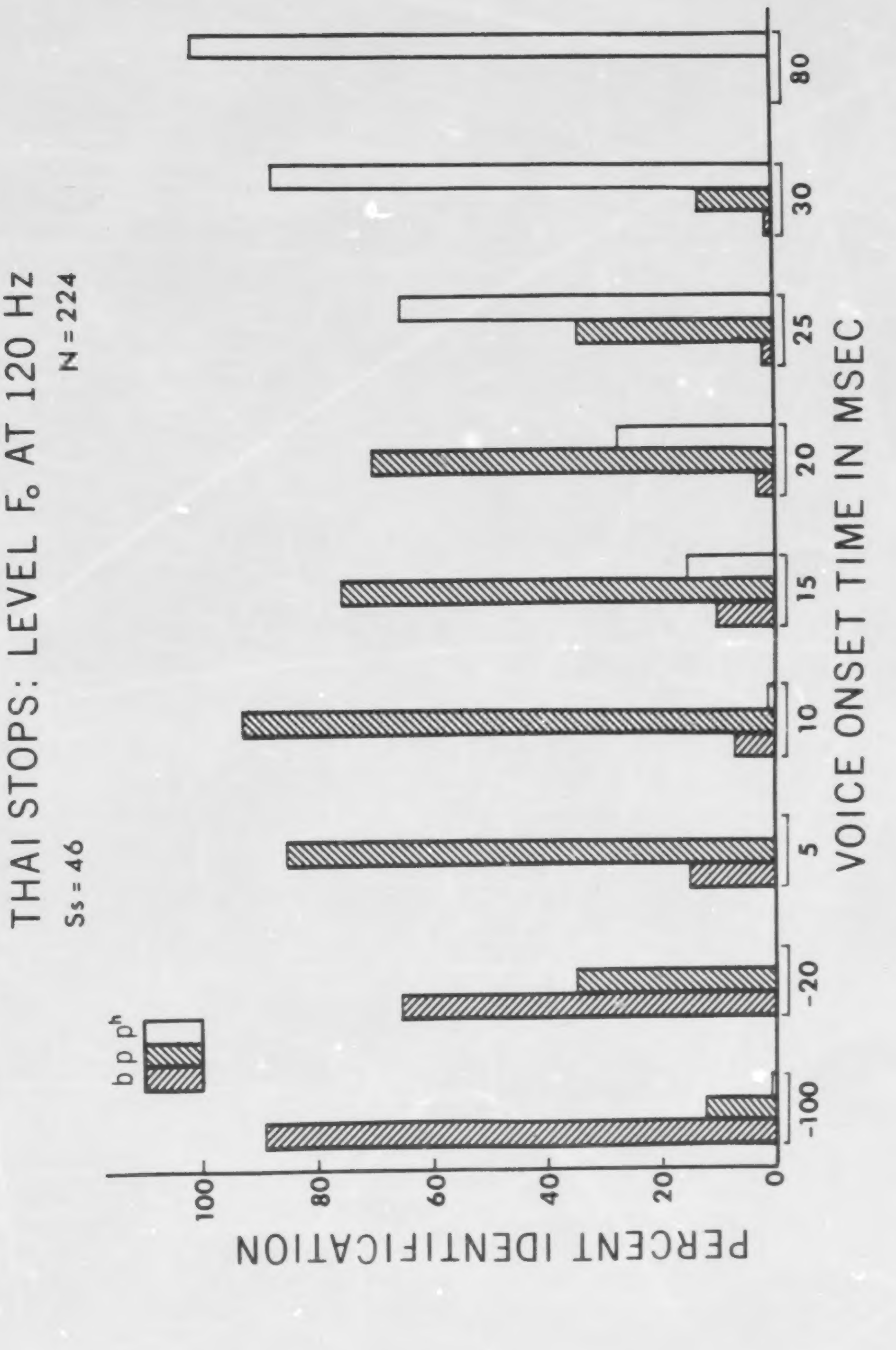

THAI STOPS: LEVEL F_0 AT 120 Hz
Ss = 46
N = 224
b p pʰ
PERCENT IDENTIFICATION
0
20
40
60
80
100
-100
-20
5
10
15
20
25
30
80
VOICE ONSET TIME IN MSEC

FIGURE 2

THAI STOPS: 50 MSEC SHIFT OF F_0 TO 120Hz

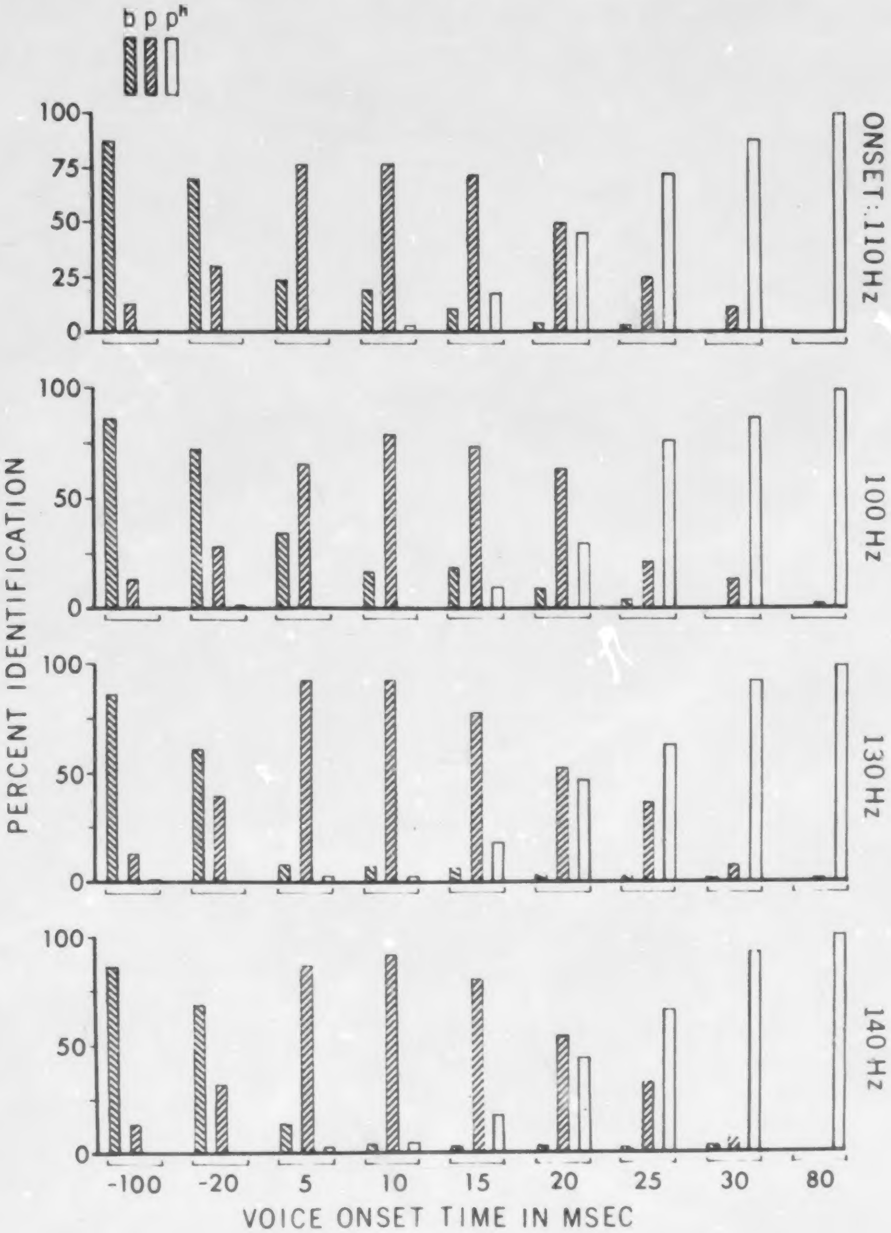


FIGURE 3

THAI STOPS: 100 MSEC SHIFT OF F₀ TO 120 Hz

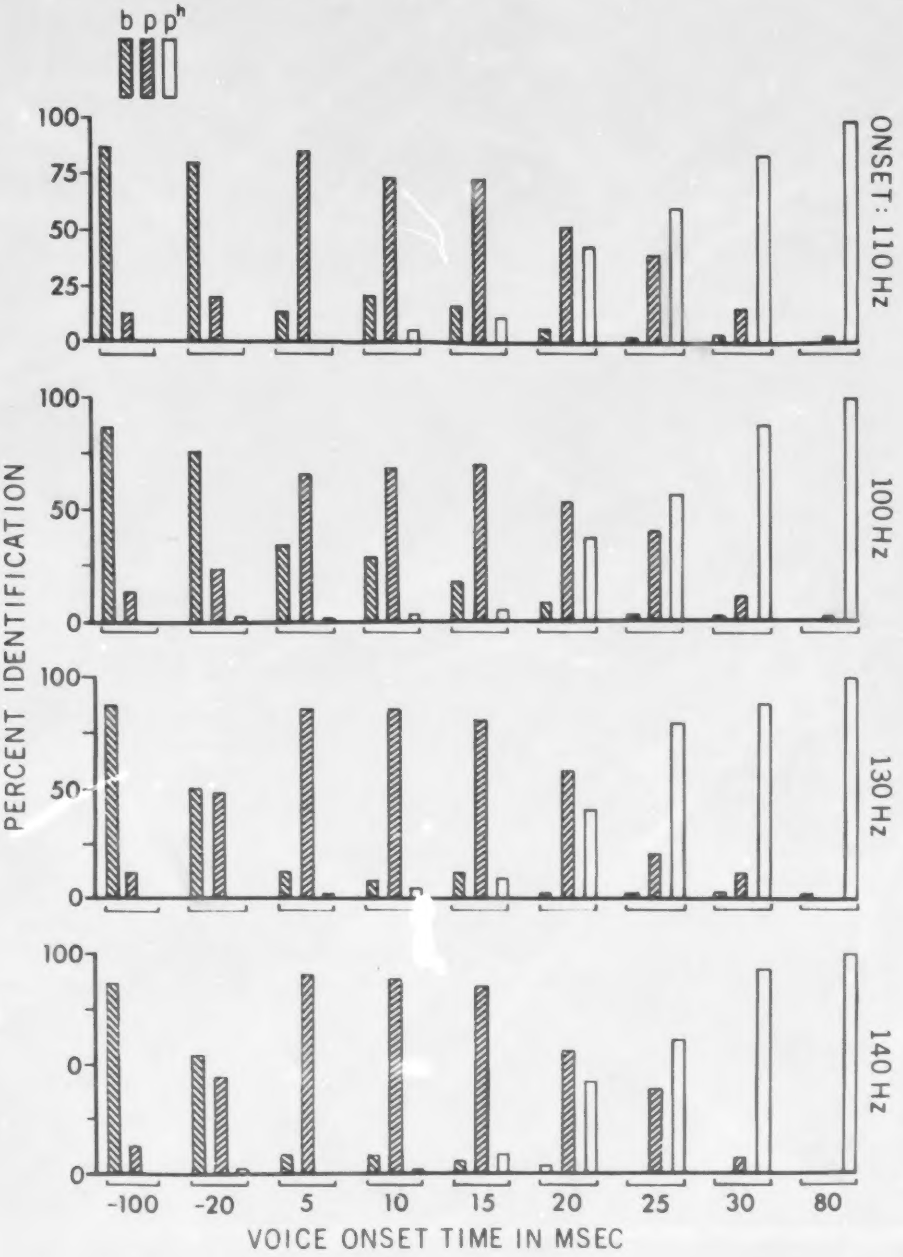


FIGURE 4

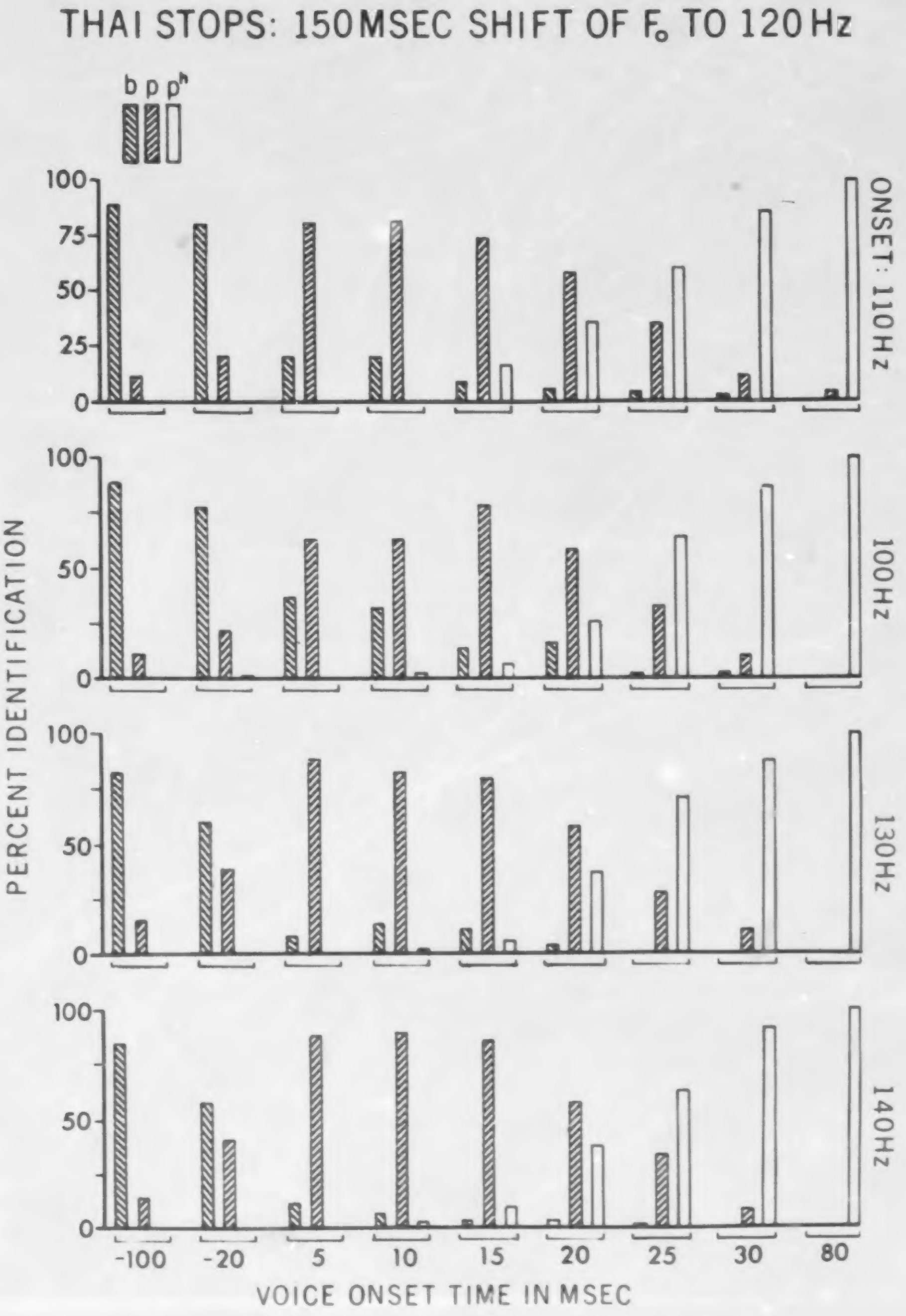

THAI STOPS: 150MSEC SHIFT OF F_0 TO 120Hz
b p pʰ
ONSET: 110Hz
100Hz
130Hz
140Hz
PERCENT IDENTIFICATION
100
75
50
25
0
-100
-20
5
10
15
20
25
30
80
VOICE ONSET TIME IN MSEC

FIGURE 5

If we look only at the two ends of the voice onset time continuum, -100 and +80 msec, we cannot see that the Thai subjects are influenced by pitch perturbations of any duration they may hear. As a matter of fact, as we look over all the bar graphs, it becomes clear that voice timing is a far more powerful cue than pitch shifts. In general, the perceptual crossover points between categories are not moved; instead, the distribution of values within each category is pushed and pulled in both directions. If, however, you scan down the -20-msec columns in Figure 4 for 100 msec and Figure 5 for 150 msec, you will find that for onset values of 130 and 140 Hz we do have a boundary shift; for the distinction between the voiced and voiceless unaspirated stops, the boundary shifts leftward from -7 msec to about -20 msec with more stimuli assigned to the voiceless category under the influence of a long duration of fundamental frequency fall. Although the 50-msec shift seems to be too short to provide for a boundary change, at +5 msec we see the voiced category succumbing to the voiceless inaspirate. The boundary between the voiceless unaspirated and aspirated stops does not shift at all, remaining at about +22 msec. Indeed in the region of this boundary it is hard to see a consistent trend. One exception would appear to be on the 50-msec display where we see that at this boundary for the 100-Hz onset, responses are pulled from the aspirate to the inaspirate, as might be expected. A refined statistical analysis, yet to be performed, may yield a few more subtle tendencies.

We may conclude then that perturbations of fundamental frequency at the beginnings of syllables with initial stops can influence voicing judgments in Thai. The effect is enhanced with greater durations of frequency shift. It seems to favor the boundary region between the voiced stop and the voiceless inaspirate. There are some effects at the boundary between the two voiceless categories, but they are less consistent and not easy to interpret. Shifts of fundamental frequency, to be ascribed, along with the feature of voice onset time, to states of the larynx have some cue value in Thai, although they are clearly subordinate to voice onset time. That is, the effects are most marked in zones of perceptual ambiguity along the voice timing continuum.

The effects found in this study do lend some support to the argument that the emergence of tones in Proto-Tai or, perhaps, the increase in the number of tones, could have been a conditioning factor in the shifting and merging of consonant voicing categories. We can imagine something like the following situation. As the pitch perturbations associated with the voicing states of initial consonants became apparent to speakers of the language and gradually moved toward phonemic status as tones, the vowel allophones with their concomitant pitch characteristics became more and more differentiable. The pitch coloring must have taken up increasing amounts of time as it became more noticeable; this is implied by my effect with the longer durations of fundamental frequency shift. Brown (1965) has suggested that speakers of the language concentrated perceptually more on the central portion of the word at the expense of attention to the initial consonant when arriving at a lexical decision. In this way, the syllable initial became less and less important. We may speculate that children learning these lexical classes with a shifted perceptual set must have begun to rearticulate the initials, as a deviation from the practice of their elders, more in conformity with what they heard, namely, a shift in the voicing boundary conditioned by pitch. Finally, my data make the unaspirated stop at least as reasonable as an intermediate stage for the change from voiced to voiceless aspirated stop in modern Thai as is the murmured stop posited by Haudricourt and Martinet (1946).

REFERENCES

Abramson, A. S. and L. Lisker. (1965) Voice onset time in stop consonants: Acoustic analysis and synthesis. In *Proceedings of the 5th International Congress of Acoustics*. (Liège: Imp. G. Thone) A51.

Brown, J. M. (1965) *From Ancient Thai to Modern Dialects*. (Bangkok: Social Science Association Press).

Coedès, G. (1949) Les langues de l'Indochine. Conférences de L'Institut de Linguistique de l'Université de Paris 7, 63-81.

Erickson, D. (in press) Phonetic implications for an historical account of tonogenesis in Thai. In *Studies in Tai Linguistics*, ed. by J. Chamberlain and J. G. Harris. (Bangkok: Central Institute of English Language).

Fujimura, O. (1971) Remarks on stop consonants: Synthesis experiments and acoustic cues. In *Form and Substance*, ed. by L. L. Hammerich, R. Jakobson, and E. Zwirner. (Copenhagen: Akademisk Forlag), pp. 221-232.

Gandour, J. (1974) Consonant types and tone in Siamese. J. Phonetics 2, 337-350.

Gedney, W. J. (1974) Future directions in comparative Tai linguistics. Unpublished manuscript.

Haggard, M., S. Ambler, and M. Callow. (1970) Pitch as a voicing cue. J. Acoust. Soc. Amer. 47, 613-617.

Haudricourt, A.-G. and A. Martinet. (1946) Propagation phonétique or évolution phonologique? Assourdissement et sonorisation d'occlusives dans l'Asie du sud-est. Bulletin de la Société de Linguistique de Paris 43, 82-92.

House, A. S. and C. Fairbanks. (1953) The influence of consonant environment upon the secondary acoustical characteristics of vowels. J. Acoust Soc. Amer. 25, 105-113.

Lea, W. A. (1973) Segmental and suprasegmental influences on fundamental frequency contours. In *Consonant Types and Tone*, ed. by L. M. Hyman (Southern California Occasional Papers in Linguistics No. 1), pp. 15-70.

Lehiste, I. and G. E. Peterson. (1961) Some basic considerations in the analysis of intonation. J. Acoust. Soc. Amer. 33, 419-425.

Li, Fang-Kuei. (1947) The hypothesis of a pre-glottalized series of consonants in primitive Tai. Academia Sinica 11, 177-188.

Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word 20, 384-422.

Maspéro, H. (1911) Contribution à l'étude du système phonétique des langues thai. Bulletin de l'Ecole française d'Extrême-Orient 19, 152-169.

Facial Muscle Activity in the Production of Swedish Vowels: An Electromyographic Study

Katherine S. Harris,[*] Hajime Hirose,[+] and Kerstin Hadding[++]

INTRODUCTION

The purpose of this paper is to specify further the nature of vowel rounding in Swedish, using electromyography (EMG) to clarify the role of several facial muscles.

Swedish is conventionally analyzed as having 18 vowel phonemes in stressed position, nine of which are long and nine short.[1] Alternatively, the vowel system is analyzed as consisting of nine qualitatively different vowel pairs, each with one long and one short member. (See, among others, Malmberg, 1956; Elert, 1964, 1970; Öhman, 1966.) In spite of the qualitative difference that exists in a varying degree between long vowels and their short counterparts, duration may be considered the relevant feature in Swedish (for discussion, see Hadding-Koch and Abramson, 1964). In some recent studies only consonant length is specified in the underlying forms (Teleman, 1969; Linell, 1973), vowel length being derived from the morpheme structure (Eliasson and La Pelle, 1970).

For the purpose of the present investigation, 18 vowels may be listed. They are given below in International Phonetic Alphabet (IPA) transcription, with a few exceptions, together with Swedish key words and examples in English, German, or French, depending on the presence of vowels of similar quality. In addition, the symbol [:] is used to indicate the long member of a pair.

---

[*] Haskins Laboratories, New Haven, Conn., and the Graduate Division of the City University of New York.

[+] Faculty of Medicine, University of Tokyo.

[++] Department of Linguistics, University of Lund.

[1] In most Swedish dialects the distinction is no longer upheld between short /e/ and short /ɛ/, which are both pronounced [ɛ], yielding a system of nine long vowels and eight short ones.

Acknowledgment: Frances Ingemann has been extremely helpful, both in supplying information about Skandinavian vowel systems, and in considering questions of interpretation.

[HASKINS LABORATORIES: Status Report on Speech Research SR-41 (1975)]

Symbols	Key words		Symbols	Key words	
	Swedish	English, German, French		Swedish	English, German, French
[i:]	rita	beat (E), bieten (G)	[I]	ritt[2]	bit (E), bitten (G)
[e:]	reta	beten (G)	[e]	rett	Bett (G)
[ε:]	räta	bäten (G)	[ε]	rätt	bed (E)
[y:]	ryta	Fühlung (G)	[Y]	rytt	Fülle (G)
[ø:]	rota	Höhle (G)	[œ]	rott	Hölle (G)
[ɯ:][3]	ruta	----------	[θ][3]	rutt	----------
[u:]	röta	boot (E), fou (F)	[U][4]	rött	foot (E), foule (F)
[o:]	Rota	holen (G)	[ɔ]	rått	Holle (G)
[α:]	rata	far (E), pâte (F)	[a]	ratt	patte (F)

AIM OF STUDY

The aim of the investigation was primarily to study the rounding feature--11 of the Swedish vowels are assumed to be more or less rounded--and to compare the muscle activity involved in the production of rounded, spread, and neutral vowels.

Swedish rounded vowels are particularly interesting because more than one type of rounding has been suggested, as noted by early Swedish phoneticians (e.g., Lyttkens and Wulff, 1885; Noreen, 1902-07; Danell, 1911). Although their descriptions of the rounded vowels vary, they agree that [y] is articulated with protrusion of the lips and marked labialization (Noreen), lips protruded and outrounded, almost tubelike (Lyttkens and Wulff). On the other hand, [ɯ] is said to be narrowly rounded, the lips being "indrawn" rather than protruded; while [u] is described as narrowly rounded with slightly protruded lips. It is clear that protrusion and rounding decrease as the mouth opening increases. The vowel [α] is thus said to be "possibly somewhat labialized" (Danell, 1911:37), and "broadly labialized (and with somewhat protruded and also somewhat rounded lips)" (Noreen, 1903-07:529).

Malmberg (1956:317), describing the vowel system at a higher level of abstraction, states:

---

[2] In Swedish, short vowels are followed by a long consonant and long vowels are followed by a short consonant, as indicated by the orthography.

[3] The symbol [ɯ] is taken from the Swedish dialectal alphabet. The vowel can be described with reference to the IPA cardinal vowel [ʉ], which is a high central rounded vowel. The Swedish vowel is more fronted and less high. [θ] is a central half-open vowel, less rounded than [ɯ].

[4] The symbol [U] is used in the Swedish dialectal alphabet. The corresponding IPA symbol is [ω̂].

173

> Already some early Swedish phoneticians...had pointed to a difference in lip articulation which was supposed to characterize [ɯː] as opposed to [y:] and [ø:]. And later investigations have proved that there is a clearcut difference in lip closure between the [y]-[ø]-[œ] series on one hand, and the [ɯ]-[θ] series on the other. The lip opening is smaller for the latter type and there is no protrusion of the lips, only a strong closure. Consequently, the mouth cavity resonance is lowered by this smaller opening.[5]

In a recent personal communication, Malmberg has explained that by protrusion of the lips he meant the outrounded variety, most clearly represented by [y:], and that "protrusion," differently defined, may well be present also in other rounded vowels. For the second type of rounding, "puckering" or "pursing" may be a better description.

Fant (1971:260), who like Malmberg describes the vowel system in terms of distinctive features, states:

> The [ɯ:] can have the same degree of tongue height as [ø:] whilst the phonetically distinctive element of [ɯ:] is an extreme narrowing of the lips, which generally is realized as a diphthongal transition to lip closure and back to a more open terminal phase. This feature [ɯ:] shares with [u:]. They are traditionally referred to as being "inrounded" whilst the [y:], [ø:] and [o:] have a lesser degree of lip narrowing and are said to be "outrounded," referring to the protrusion of the lips. A diphthongal movement towards articulatory closure and back to a more open phase is also typical of long [i:] and [y:]. This is a matter of tongue body movement, whereas it is not always recognized that the main element of the [u:] and the [ɯ:] diphthongs is a lip closing gesture.[6]

Thus, we may ask several questions about the lip muscle action underlying rounding in long vowels and their short counterparts.

Is the difference between vowels merely a matter of the relative intensity of muscle activity, or are there differences in the pattern of activity of the various muscles around the lips? Is the timing of the patterns different for the different vowels?

PROCEDURE

Kerstin Hadding served as subject for this experiment. She speaks a southern Swedish variant of standard Swedish.[7] The experiment was repeated three times within a six-week period, with some variations between runs, as described in Table 1.

---

[5] Malmberg's symbol [ŭ] has been changed to [θ] to conform to usage in this paper.

[6] Fant's symbol [ʉ] has been changed to [ɯ] to conform to usage in this paper.

[7] The dialect is similar to that of Malmberg.

TABLE 1: Utterances in spoken samples.

RUN I

"to say ______ again"

"to say ______"

V = [i:], [y:], [ɯ:], [u:], [ɑ:], [I], [Y], [a].

Numbers of tokens of each vowel: 20.

RUN II

"to say ______"

V = [i:], [e:], [ɛ], [y:], [o:], [ø:], [u:],
[o:], [ɑ:], [I], [ɛ], [Y], [œ], [θ], [U],
[ɔ], [a].

Numbers of tokens of each vowel: 20.

RUN III

V = [i:], [y:], [ɯ:], [u], [ɑ:], [I], [Y],
[θ], [U], [a].

Numbers of tokens of each vowel: 16.

Hooked-wire electrodes, similar to those described by Hirano and Ohala (1969) were used in the present experiment. Detailed notes on electrode preparation and insertion technique are given by Hirose (1971). The placements are similar to those described by Leanderson, Persson, and Öhman (1971), except for those in the buccinator (BUC) and the orbicularis oris at the angle of the mouth (OOA), which were not included in their study. Insertion to BUC was made approximately 2 cm lateral to the angle of the mouth, superficially enough to place the electrodes in its thin muscle layers. The OOA is reached at the angle of the mouth on the vermillion border.

In the three runs, successful recordings were obtained from various facial muscles as given in Table 2.

For verification of the correct placement of the electrodes, the subject was required to attempt various articulatory as well as nonarticulatory gestures, which were assumed to involve the muscles to be studied. The electrodes at the angle of the mouth might be suspected of contamination by BUC or depressor anguli oris (DAO), or even by other muscles not included in the present investigation, e.g., the levator group. However, since the activity recorded at the angle of the mouth was similar to that recorded by the other two OO electrodes but did not coincide with that of BUC or DAO, it was assumed to represent the activity of that particular portion of OO alone.

All the EMG signals were recorded on a multichannel data recorder simultaneously with acoustic signals and timing markers. The signals were then reproduced and fed into a computer after appropriate rectification and integration. The EMG signals from each electrode pair were averaged over about 15 selected

TABLE 2: Muscles inserted in each run.

Muscles	Runs I	Runs II	Runs III
Orbicularis oris superior (OOS)	0	0	0
Orbicularis oris inferior (OOI)	0	0	0
Orbicularis oris at the angle of mouth (OOA)	0	0	0
Buccinator (BUC)	X	0	0
Depressor labii inferioris (DLI)		0	0
Depressor anguli oris (DAO)	0		0
Mentalis (MENT)		X	
Anterior belly of the digastric (AD)*	0	0	0

0 = record obtained, X = recording attempted but failed, * = anterior belly of the digastric (will not be discussed in this paper).

utterances of each test word, with reference to a lineup point on the time axis representing a predetermined acoustic event in the speech signal. In the present study, the onset of voicing of the stressed vowel in the test word was chosen for the lineup. The data recording and computer-processing systems used in the present experiments are described in more detail by Port (1971, 1973).

RESULTS AND DISCUSSION

In the present study, simultaneous mapping of the activity of several muscles was possible.

As might be expected from the literature, lip rounding and spreading are clearly differentiated by the various electrodes (Harris, Lysaught, and Schvey, 1965; Fromkin, 1966; Tatham and Morton, 1969; Hadding, Hirano, and Smith, 1970; Leanderson et al., 1971; Leanderson and Lindblom, 1972). Owing to the use of two consonant frames and a number of vowels, a number of potential interaction patterns between consonant and vowel may be examined. Figure 1 illustrates this point. The results show continuous activity when the muscle is active for both consonant and vowel (D:OOS), or reciprocal activity and suppression, when the muscle actions are antagonistic (C:OOS vs. BUC). In cases where the consonant is neutral (as in B:OOS, A:BUC), there will be anticipatory coarticulation (Daniloff and Moll, 1968; Lubker, McAllister, and Carlson, 1974).

[d] Context

There are always activities for rounding on the electrodes OOS, orbicularis oris inferior (OOI), OOA, and depressor labii inferioris (DLI); however, small differences appeared between runs, and between electrodes assumed to be in the fibers of the same muscle, as discussed below.

Results for the three long rounded vowels [y:], [ɯ:], and [u:] for three runs, are shown in Figure 2. Curves are similar for OOA on the three runs, in

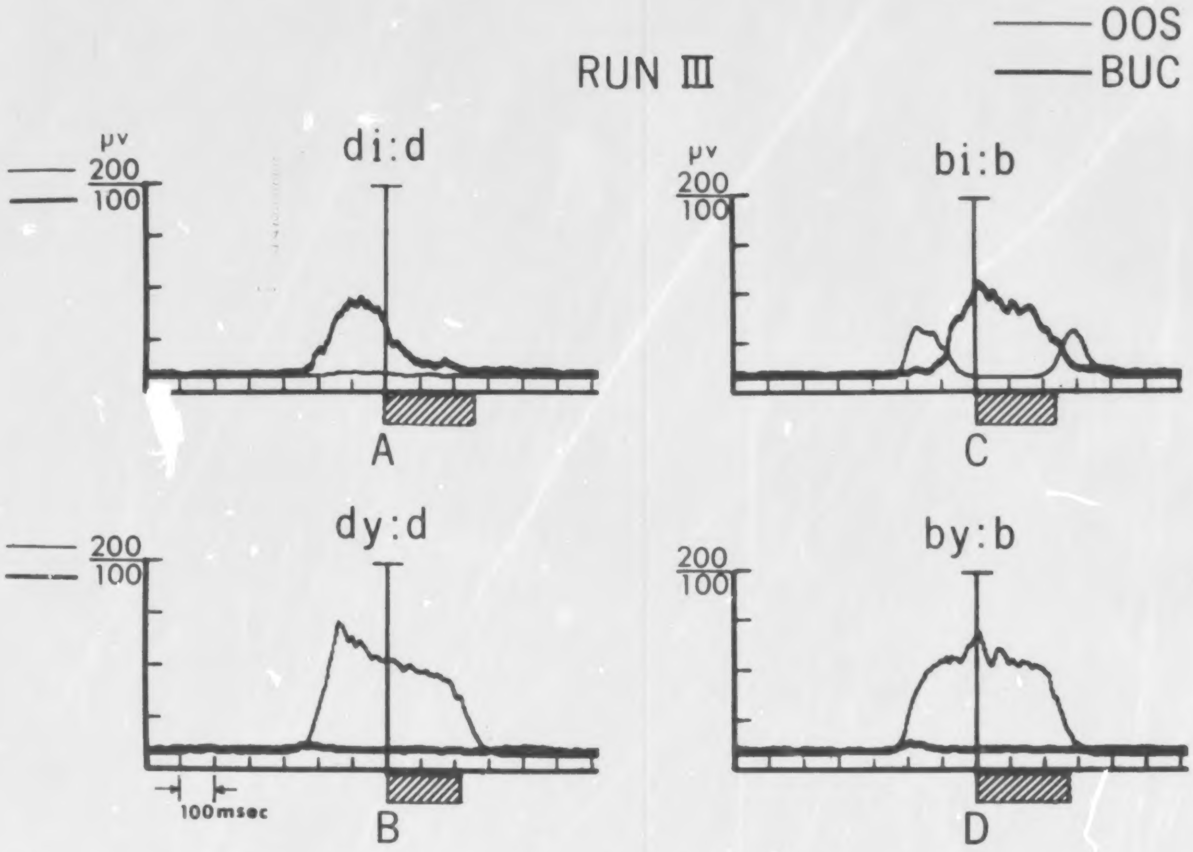


Figure 1: Superimposed, averaged EMG curves of orbicularis oris superioris (OOS) and buccinator (BUC) muscles for four utterances: [ha'di:da], [ha'dy:da], [ha'bi:ba], and [ha'by:ba]. The lineup for averaging is the stressed vowel onset, indicated by the vertical line. The duration of the stressed syllable is indicated by the shaded bar beneath each figure.

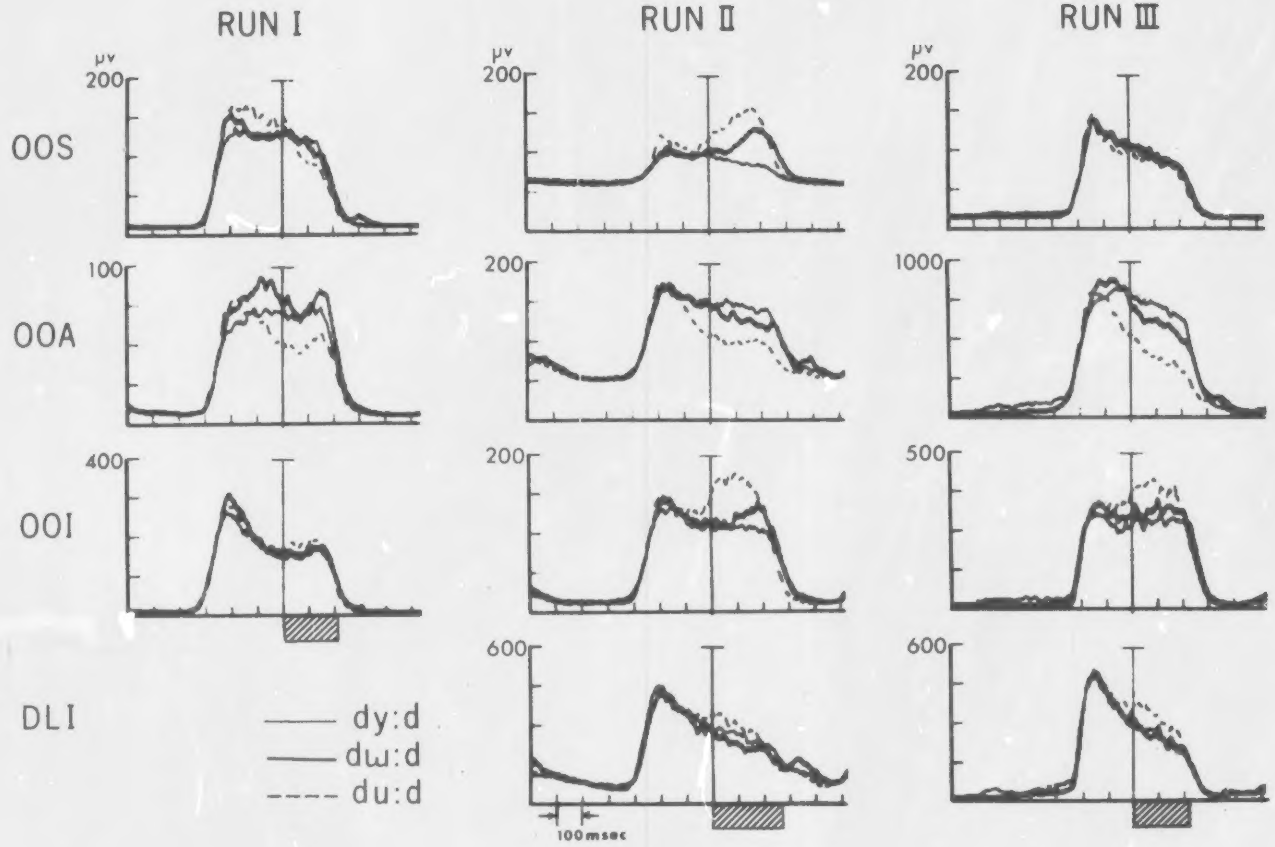


Figure 2: Averaged EMG curves for four muscle placements, superimposed for utterances containing the vowels [y:], [ɯ:], and [u:], in three runs. The approximate duration of the stressed syllable is indicated by the shaded bar beneath each column.

that small and inconsistent differences are seen for [y:] and [ɯ:]; while there is a rapid decline in the early activity for [u:]. The other lip electrodes, OOS and OOI, show two types of pattern--either the three vowels show little or no difference (OOS Runs I, III; OOI Run I), or there is late activity for [u:] and [ɯ:] (OOS Run II; OOI Runs II, III). The vowel [ɯ:] is intermediate between [u:] and [y:]. The run-to-run differences do not depend, apparently, on electrode sensitivity. For example, in Run I the scale factor for OOI is 400 μv, but the three vowels are undifferentiated, while on Run II, the scale factor is only 200 μv, but the three curves are quite separate. A reasonable hypothesis is that the different insertions were picking up different fiber types, or fiber-type mixtures, and that some of the fibers were active for a late vowel component.

The DLI too was active for rounding in this subject, and showed slightly more late activity for [u:]. In one previous study, DLI was reciprocal to the OO group (Leanderson et al., 1971) but in another (Leanderson and Lindblom, 1972), DLI apparently showed patterns synergistic with the OO group. DLI may be active for rounding in supporting the soft tissues around the OO group, or DLI activity may simply represent cocontraction. (Compare also Hadding et al., 1970:7.)

Data for the rounded back vowel [o:] and rounded front vowel [ø:] are shown in Figure 3. Since these vowels were examined only on Run II, patterns should be compared only with the middle row on Figure 2. The amplitude of the early component is roughly comparable to that for [y:] and [u:] for all four electrodes. Results for the late component show [o:] to be more similar to [u:] than to [y:] in showing increased late activity for OOS, OOI, and DLI and decreased late activity for OOA. On the other hand, [ø:] is more similar to [y:]. Referring back to the quotations from Fant and Malmberg at the beginning of the paper, we note that both these vowels are grouped with [y:]. Therefore, our results support Fant's descriptions for [ø:] but not for [o:].

Results for the long vowel [ɑ:] are shown in Figure 4, with its short counterpart [a]. It may be noted that [a] does not show any consistent evidence of rounding. This point will be discussed further below, in connection with the short vowels. The long vowel [ɑ:] showed no consistent evidence of rounding on the upper lip; the pattern of rounding was therefore quite different than for any other rounded vowel. Patterns for the other three electrodes were similar to [o:].

The location DAO shows no indication of activity for any long vowel in the [d] frame. Of course, [o:] and [ø:] could not be checked, since these vowels were not part of the corpus, except on Run II, when DAO locations were not observed. The location BUC shows, as expected, no activity for any vowel normally described as rounded. It shows activity only for the spread vowel [i:], as shown in Figure 1.

[b] Context

The data for the five long vowels examined in [b] context are shown in Figure 5, for the four "rounding" electrodes. Closure peaks can be seen for [b] where the vowel is not rounded or not fully rounded, as in [ɑ:] and [i:]. When the vowel is rounded, the situation is more complicated. For OOS and OOA there

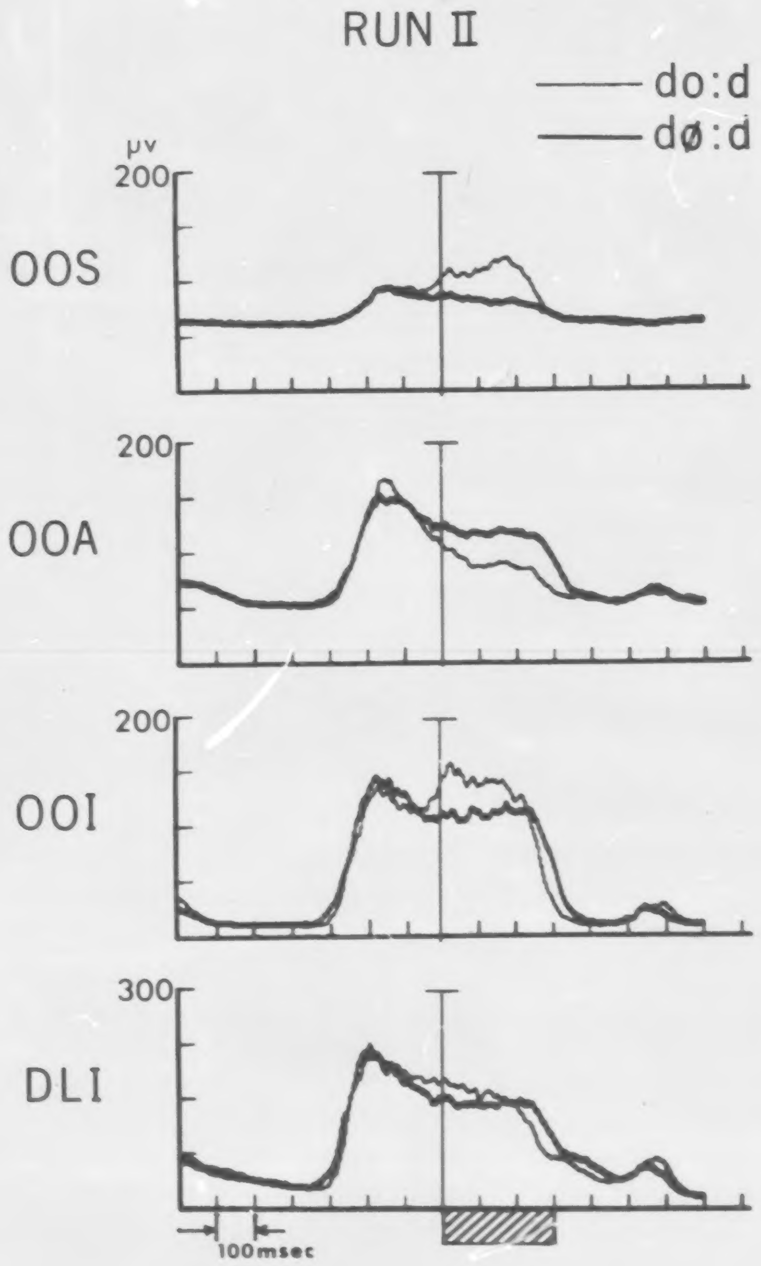


Figure 3: Averaged EMG curves for utterances containing [dø:d] and [do:d].

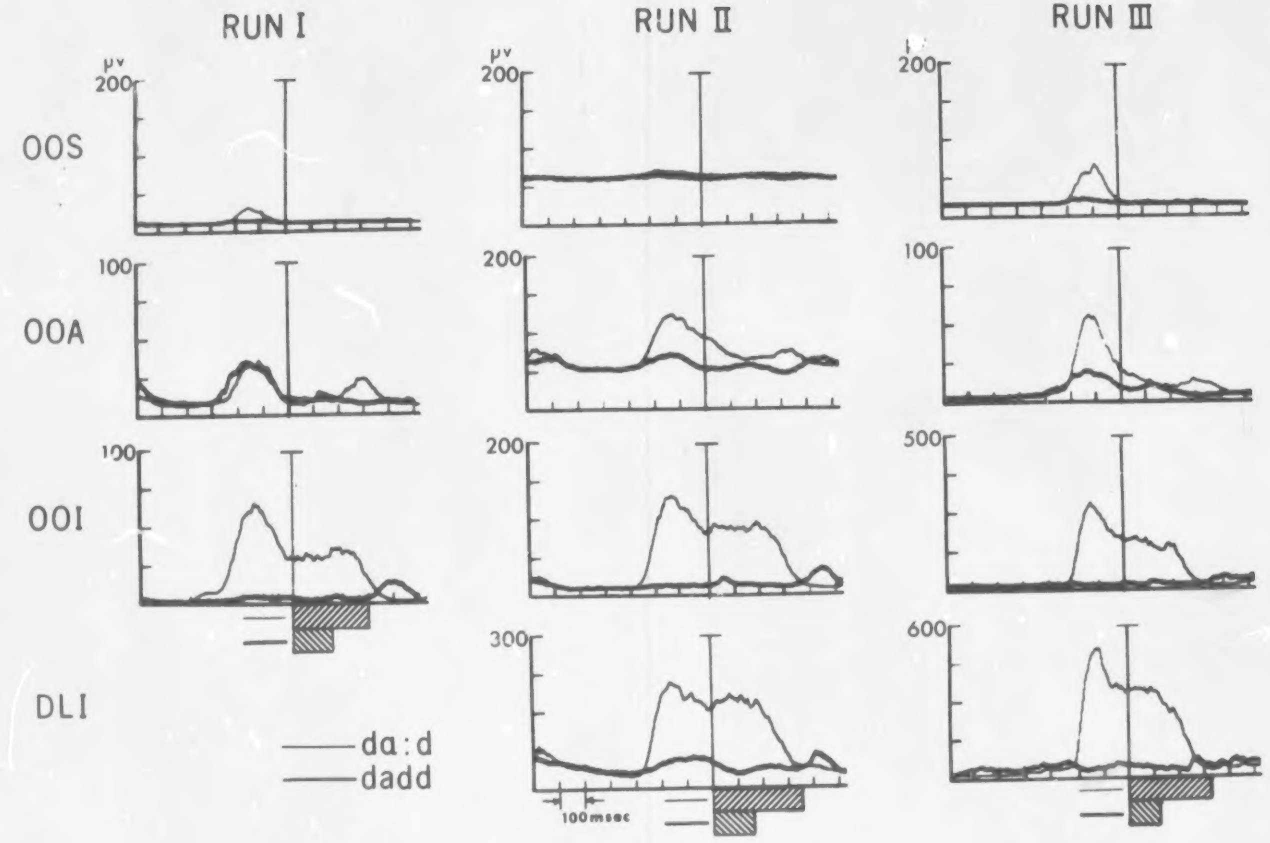


Figure 4: Averaged EMG curves for utterances containing the long vowel [ɑ:] and the short vowel [a]. The approximate duration of syllables containing long and short vowels are shown by the shaded bars.

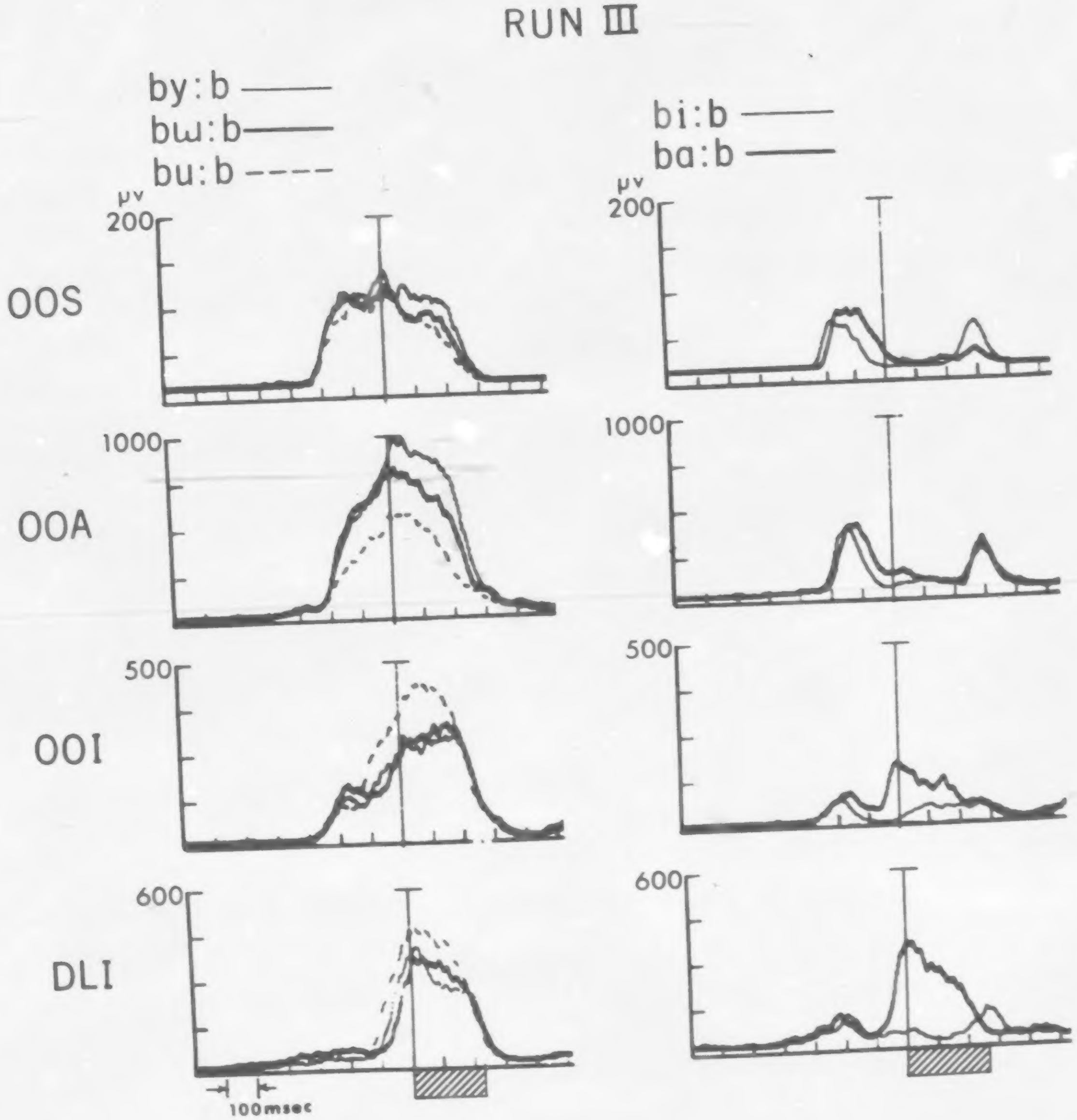


Figure 5: Averaged EMG curves for utterances containing five long vowels in [bV̄b] frame, for four electrode positions. Approximate durations of syllables are shown by the shaded bars.

seems to be a merging of the activity for consonant and vowel. It is interesting to note that this results in less activity before the lineup than in [d] context, where anticipatory coarticulation occurs. For OOI, there is a [b] closure peak, followed by suppression, apparently associated with consonant release, followed by rounding for the vowel. For DLI, apparently, lip compression for [b] is incompatible with vowel activity; there is no evidence of a consonant compression peak, and vowel activity begins later. Although the result might suggest some sort of anticipatory suppression of [b] closure, the [b] peak for the vowel in the [i]-[a] context is very small, so the context difference may not be a reliable one.

Although there is a difference between [d] and [b] contexts, the relations among the late components for the three vowels [y:], [ɯ:], and [u:] are quite similar--OOS activity more-or-less the same, greater activity for OOI and DLI for [u:], and greater activity for [y:] at OOA, with [ɯ:] falling somewhere between. It is worth noting that OOI and DLI, the two electrodes whose activity patterns are incompatible with closure release, are also the electrodes showing greatest activity for [u:]; a vowel characterized as "inrounded," as is [ɯ:].

In [d] context (Figure 2), differences between the three vowels for the early component are quite similar, suggesting that the difference between the three vowels for this speaker is entirely in a diphthongization component. However, if the three vowels are produced in an identical way, they should interact in the same way with a frame change. It is clear that the early part of the gesture is not the same for [y:] as for the other two vowels, since curves for at least OOA and OOI are different.

There is no separate peak for the terminal [b]. However, the result is entirely consistent with earlier work of Bell-Berti and Harris (1973) on the mylohyoid and palatoglossus muscles. They suggest that if a muscle is active for both members of a CV sequence, separate peaks will be seen for both elements; in a VC sequence, the two peaks will merge. The reason for this effect is not clear.

Since [o:] and [ø:] were examined only in Run II, it is not possible to examine the [b] frame for these vowels. Results for [ɑ:] are shown in Figure 5. Clear initial [b] peaks are seen for all four electrode positions, as well as terminal peaks for three. Therefore, the nature of lip activity for [ɑ:] must be different from that for the other three rounded vowels.

There is DAO activity for both initial and terminal [b] peaks, as shown in Figure 6. Since this location shows the ordinary closure peak usually seen for [b], we can only assume that the fibers are involved in bringing the upper or lower lip closed. The same results are apparently shown qualitatively by Leanderson et al. (1971); perhaps there is supporting action of the soft tissue for closure. It is interesting that the closure peak is lower for the three heavily rounded vowels, [y:], [ɯ:], and [u:]. Apparently, the lip activity for these vowels is anticipated by a weaker closure gesture.

Short Vowels

The short vowels [y], [θ], [U], and [ɔ] [æ] are shown in dVd context in Figures 7 and 8, respectively; [y], [θ], [U] are shown in bVb context in

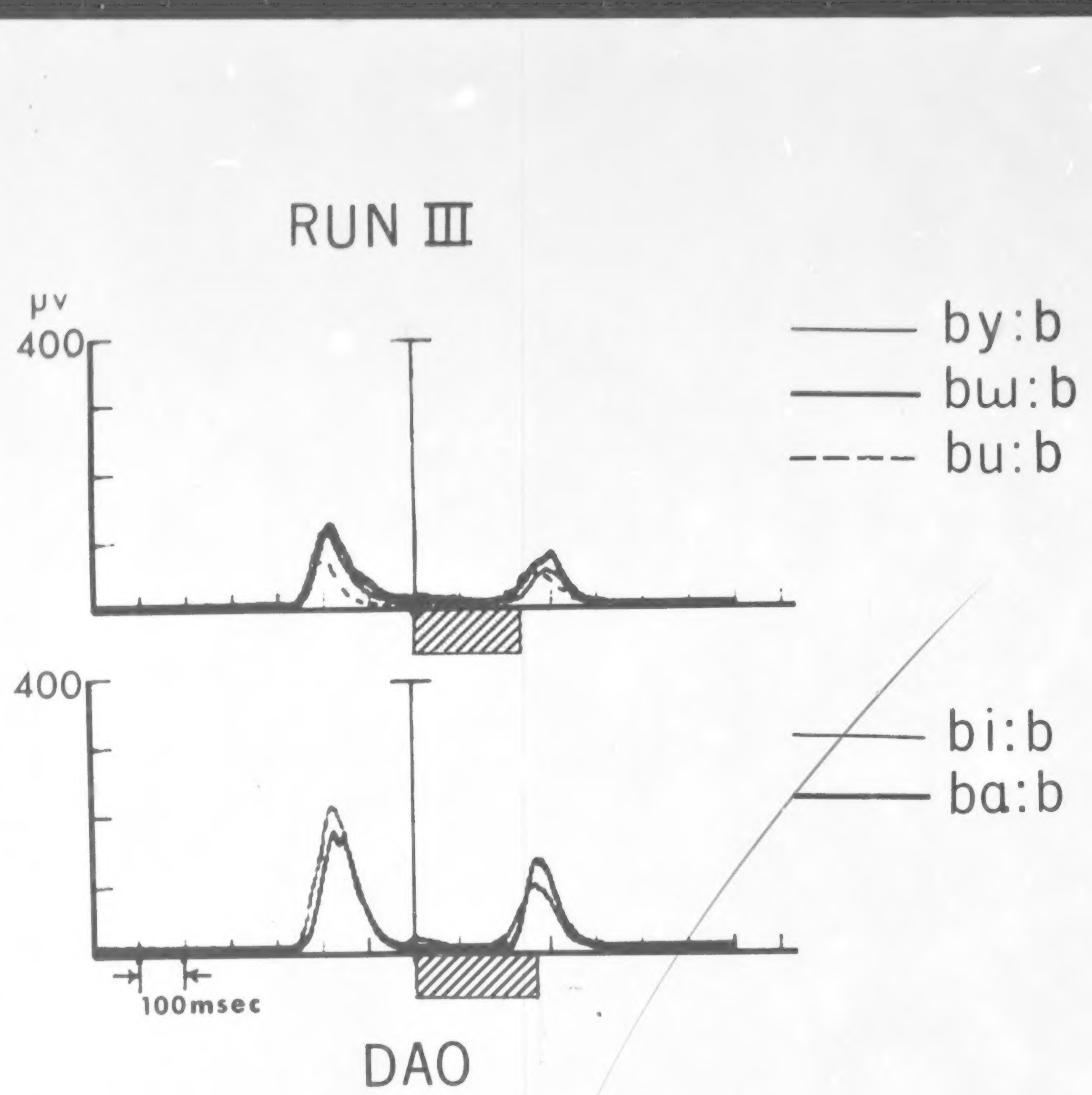


Figure 6: Averaged EMG curves for utterances containing five long vowels in [bVb] frame, for DAO electrode position. Approximate duration of syllables is shown by the shaded bars.

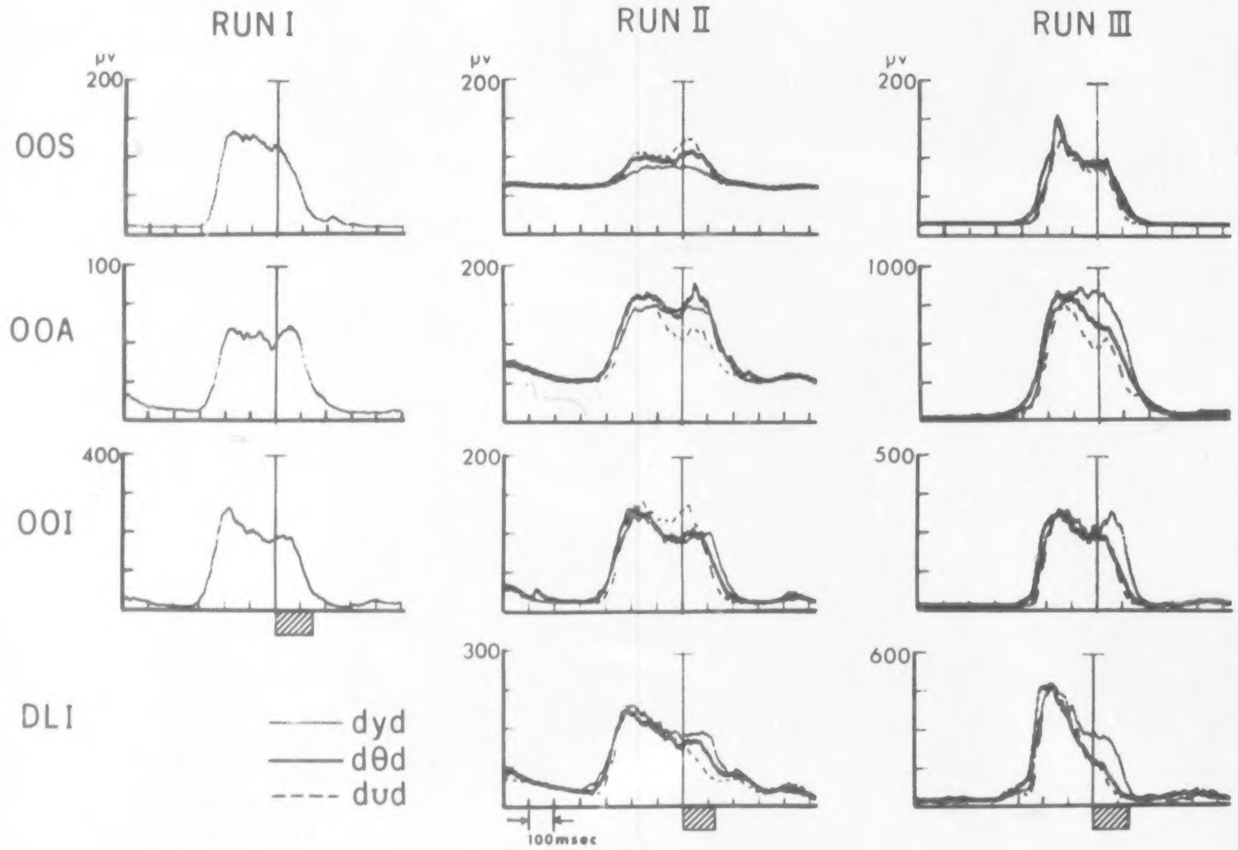


Figure 7: Averaged EMG curves for utterances containing three short vowels in [dVd] frame. Approximate duration of syllables is shown by the shaded bars.

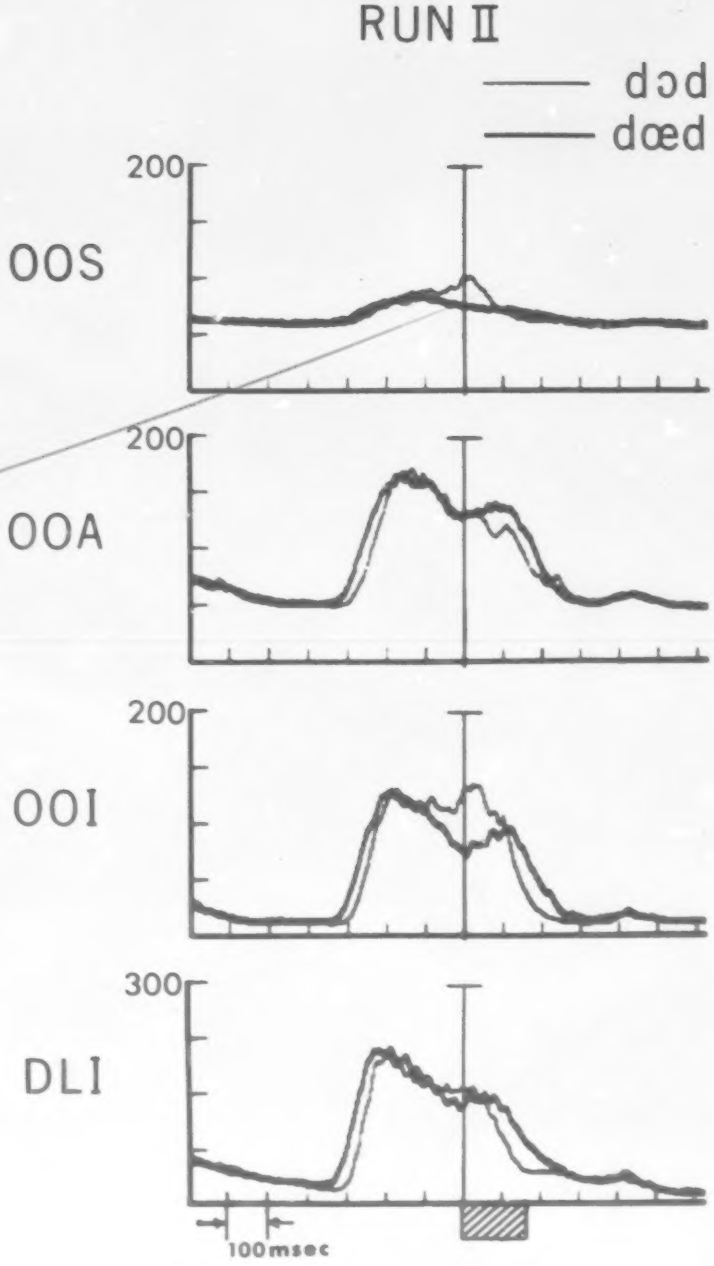


Figure 8: Averaged EMG curves for utterances containing two short vowels in [dVd] frame. Approximate duration of syllables is shown by the shaded bars.

Figure 9. The short vowel [ɑ] was shown in Figure 4. Two general differences may be noted with respect to long vowels and their short counterparts. First, differences in the late component are somewhat less consistent in the short vowels. Indeed, some of the differences in the EMG curves may occur too late to have much acoustic effect on the vowels themselves. With respect to the early component, short vowels parallel their long counterparts. The most notable difference between long and short vowels is an overall tendency for the short vowels to be produced with slightly less energy than their long counterparts. In the three runs, and two consonant frames, it is possible to compare peak height for all the electrodes consistently showing rounding activity (OOS, OOA, OOI, DLI) for all the rounded long and short pairs. Forty-seven long-short comparisons were available. In 41 of the comparisons, peak height was greater for the long member of the pair. The reversal cases were scattered among electrode positions and vowels. The overall size of the effect is not large, except for the [ɑ:-a] contrast already noted.

For all long-short pairs, a change in duration is accompanied by a change in the peak amplitude of the EMG signal. This difference should result in less extreme articulator position.

The DAO peaks are shown in Figure 10. A comparison of this figure with Figure 6 shows an interesting fact: consonant peak heights are large for both initial and terminal consonants in a short vowel environment. Furthermore, the larger peaks are somewhat longer in overall duration. Traditionally, of course, short vowels in Swedish are described as followed by a relatively long consonant, as described in the Introduction. The differences observed here in the terminal consonant are reasonable enough when viewed within this framework; however, initial consonant differences remain inexplicable. Obviously, the result must be examined in a larger corpus of material. Furthermore, an analysis, now in progress, must be completed on the accompanying acoustic signals for the vowels.

Conclusions and Discussion

Since the work described above was completed, an article on the Swedish rounded vowels has been written (McAllister, Lubker, and Carlson, 1974); we shall, therefore, summarize our own results, and compare them with theirs, as well as with earlier work on Swedish rounded vowels.

1. The traditional division of the long Swedish rounded vowels [y:], [ɯ:], and [u:] into two groups, one containing [y:] and [ø:] ("outrounded") and the other, containing [ɯ:] and [u:], was supported; however, the position of [o:] with [u:] was unexpected. These results confirm the conclusions of McAllister et al. The vowel [ɑ:], described as rounded by Elert (1964), shows characteristic rounding activity but does not group with either "inrounded" or "outrounded" vowels in the pattern of the activity.

2. All the rounded vowels (except [ɑ:]) show a pattern of rounding for locations OOS, OOI, OOA, and DLI. Patterns for OOS and OOI are very similar, as McAllister et al. remark. They did not observe OOA or DLI. Patterns from DLI are quite similar to OOI, as one might expect from its location. For three of these locations, the difference between the two vowel groups lies entirely in the

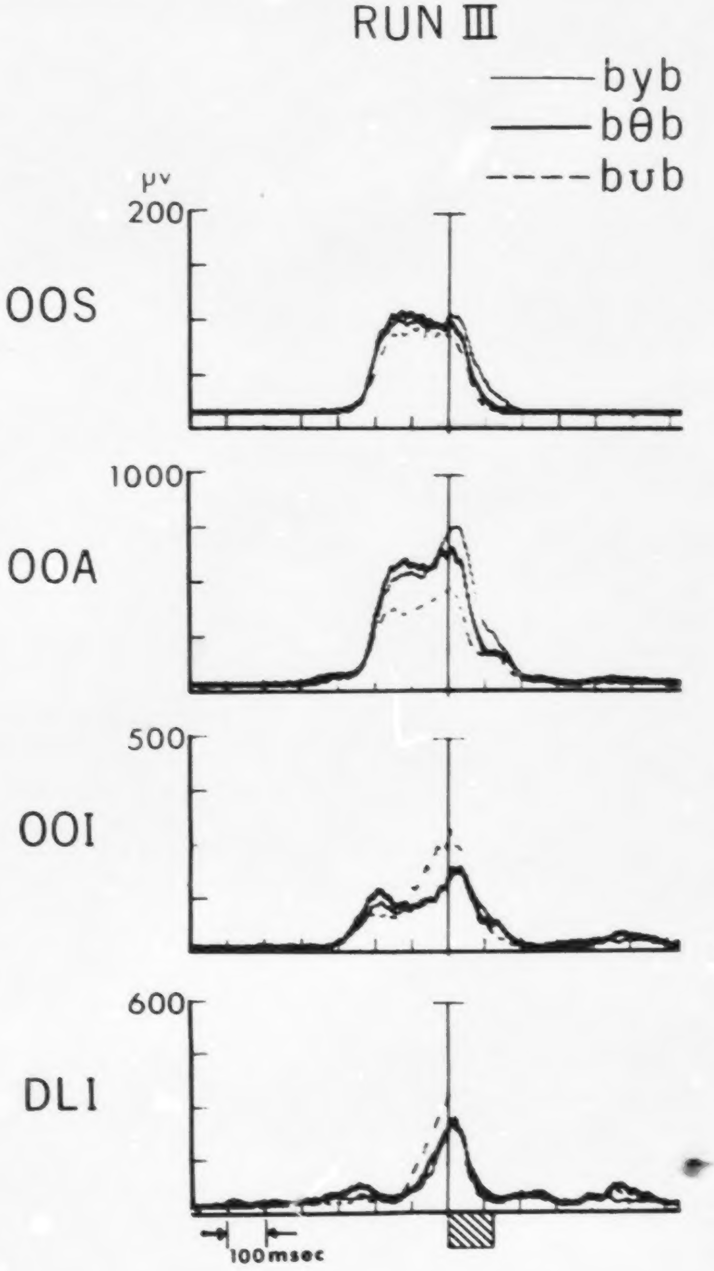


Figure 9: Averaged EMG curves for utterances containing three short vowels in [bVb] frame. Approximate durations of syllables is shown by shaded bars.

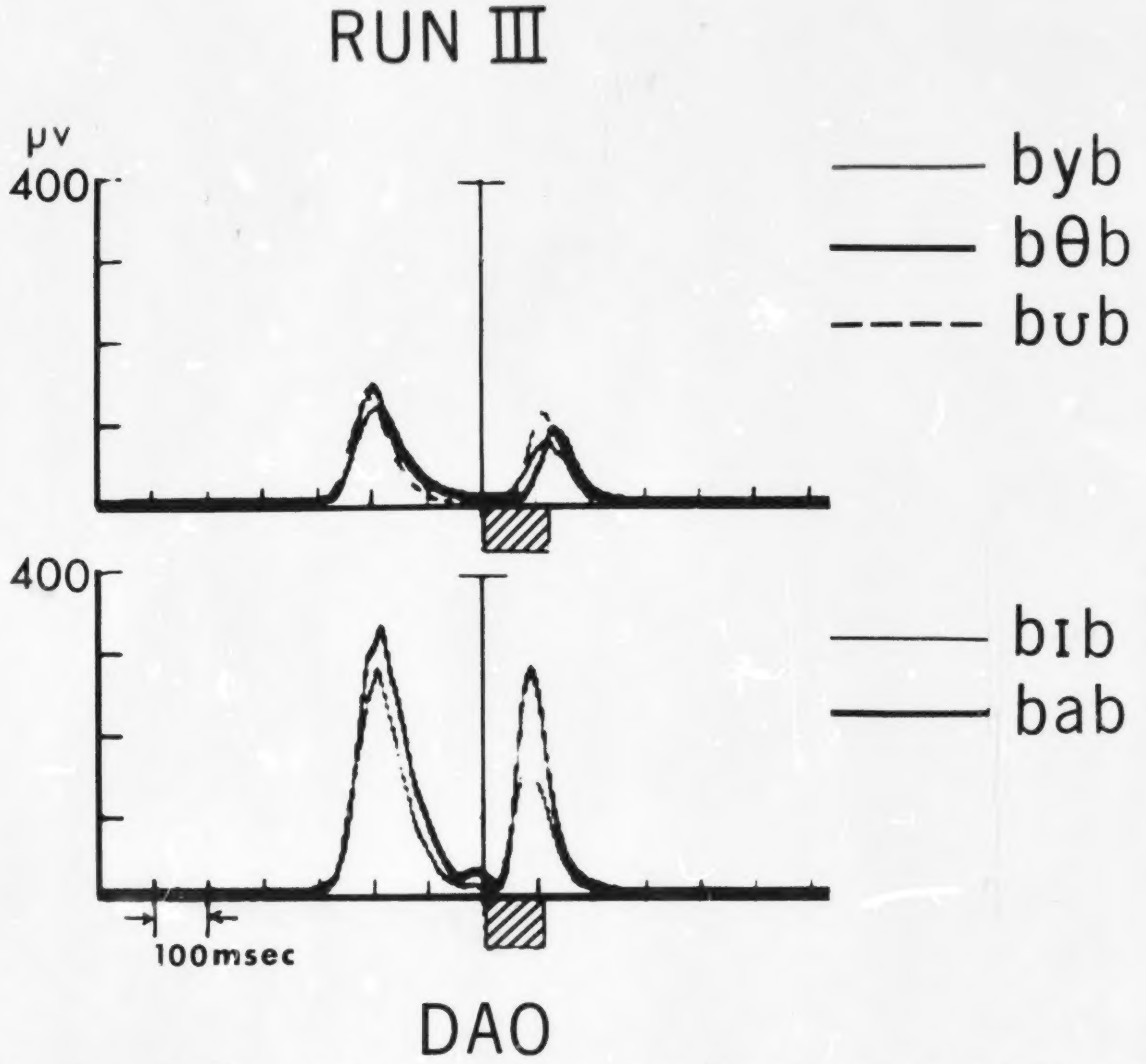


Figure 10: Averaged EMG curves for utterances containing five short vowels in [bVb] frame, for DAO electrode. Approximate duration of syllables is shown by shaded bars.

late component, for the [d] frame; the "inrounded" vowels show greater late activity. The patterns from OOA are in reverse of those from other locations; the "inrounded" vowels show less late activity. The only evidence for a difference in the early component lies in the fact that [y:] behaves a little differently in the [bVb] frame from the other vowels.

3. The location DAO shows activity for [b] closure, but not for rounded vowels (except [ɑ:]), supporting the notion that in this speaker at least, closure is either a quantitatively or qualitatively different gesture from rounding.

4. Short vowels show patterns similar to their long counterparts, but are somewhat lower in amplitude, as well as shorter in duration. The sole exception is the [ɑ:]-[a] pair, where [a] shows no consistent lip rounding. In contrast, the activity pattern for consonants surrounding short vowels is of greater amplitude.

The results of this experiment are interesting for a general theory of vowel production from two points of view.

First, they are interesting in the light they shed on the two lip-rounding descriptors--"inrounded" and "outrounded." The implication of the terms is that there is some difference in the target pattern of lip activity that pertains to the whole vowel; in fact, the differences in muscle activity pattern are quantitative rather than qualitative, and are most conspicuous in the diphthongized second part of the vowel. Evidence for difference in the early part of the lip activity pattern is indirect, at best.

The second point of general interest in these results is the relation between long and short vowel counterparts. Many languages, like Swedish, are described as having long and short vowels. In at least some of these languages, e.g., Swedish (Fant, Stålhammar, and Karlsson, 1974), Icelandic (Garnes, 1973), English (Scharf, 1964), Czech, and Serbo-Croatian (Lehiste, 1970), it can be shown that short vowels differ from their long counterparts in occupying less extreme formant target positions in an F_1-F_2 coordinate system. Two explanations might be offered for this phenomenon. One is that short vowels are "lax," relative to long "tense" vowels, and that this difference may be interpreted literally with respect to the underlying muscle activity patterns. This explanation is implicit in the Chomsky-Halle feature designation of the pairs (Chomsky and Halle, 1968), although the use of these terms is much older in the phonetics literature. This explanation requires greater activity in all relevant muscles for the tense vowel. Raphael and Bell-Berti (1975) have shown that this is not true for English. It might be argued, however, that the quantity opposition is anomalous for English, since not all vowels are paired. The explanation holds with respect to the lip muscles for the present single speaker of Swedish. Obviously, more speakers, and a larger sample of the muscles involved in the vowel articulation, must be examined.

A second explanation might be based on Lindblom's (1963) model of vowel reduction. The reduced activity of the short vowels might have something to do directly with the shorter articulatory duration. In Lindblom's model, the signals sent to the articulators for a given phone in two contexts, which result in a duration difference, are the same. However, owing to articulatory sluggishness, the shorter vowel will show greater undershoot. This explanation fails in the present case, as it does in the case of stress and speaking rate (Gay and Ushijima, 1974; Harris, 1974) because it requires a constant signal to the articulators for all conditions. Vowel signals for the longer vowel, are, in all cases examined, larger. However, it may still be that the long-short duration difference is related to an undershoot difference in some systematic way, although Fant et al. (1974:146) remark, "Swedish short vowels are not merely neutralized versions of the long vowels. The main line of contrast in any pair is along the articulatory open-close dimension, i.e., a higher F_1 for the short vowels." A detailed examination of the acoustic material from this experiment is now in progress.

REFERENCES

Bell-Berti, F. and K. S. Harris. (1973) The motor organization of some speech gestures. Haskins Laboratories Status Report on Speech Research SR-35/36, 1-5.

Chomsky, N. and M. Halle. (1968) The Sound Pattern of English. (New York: Harper & Row).

Danell, G. (1911) Svensk Ljudlära. Stockholm.

Daniloff, R. and K. Moll. (1968) Coarticulation of lip rounding. J. Speech Hearing Res. 11, 707-721.

Elert, C.-C. (1964) Phonological Studies of Quantity in Swedish. (Stockholm: Almqvist and Wiksell).

Elert, C.-C. (1970) Ljud och ord i svenskan. (Stockholm: Almqvist and Wiksell).

Eliasson, S. and N. La Pelle. (1970) Generativa regler för svenskans kvantitet. Förhandlingar vid sammankomst för att dryfta frågor rörande svenskans beskrivning VI. Umeå.

Fant, G. (1971) Notes on Swedish vowel system. In Form and Substance, ed. by L. L. Hammerich, R. Jakobson, and E. Zwirner. (Copenhagen: Akademisk Forlag), pp. 259-268. [Reprinted in Speech Sounds and Features. (Cambridge, Mass.: MIT Press (1973), pp. 192-201.]

Fant, G., U. Stålhammar, and J. Karlsson. (1974) Swedish vowels in speech material of various complexity. In Proceedings of the Speech Communication Seminar, Stockholm, 1974. (Uppsala: Almqvist and Wiksell).

Fromkin, V. A. (1966) Neuromuscular specification of linguistic units. Lang. Speech 9, 170-199.

Garnes, S. (1973) Phonetic evidence supporting a phonological analysis. J. Phonetics 1, 272-283.

Gay, T. and T. Ushijima. (1974) Effect of speaking rate on stop consonant-vowel articulation. In Proceedings of the Speech Communication Seminar, Stockholm, 1974. (Uppsala: Almqvist and Wiksell).

Hadding, K., M. Hirano, and T. Smith. (1970) Electromyographic study of lip activity in Swedish CV:C and CVC: syllables. Working Papers in Phonetics (University of California at Los Angeles) 14. [Also (1969) Working Papers (Phonetics Laboratory, Lund University) 1, 1-8.]

Hadding-Koch, K. and A. S. Abramson. (1964) Duration versus spectrum in Swedish vowels: Some perceptual experiments. Studia Linguistica 18, 94-107.

Harris, K. S. (1974) Mechanisms of duration change. In Proceedings of the Speech Communication Seminar, Stockholm, 1974. (Uppsala: Almqvist and Wiksell).

Harris, K. S., G. F. Lysaught, and M. M. Schvey. (1965) Some aspects of the production of oral and nasal labial stops. Lang. Speech *8*, 135-147.

Hirano, M. and J. Ohala. (1969) Use of hooked-wire electrodes for electromyography of the intrinsic laryngeal muscles. J. Speech Hearing Res. *12*, 362-373.

Hirose, H. (1971) Electromyography of the artiuclatory muscles: Current instrumentation and technique. Haskins Laboratories Status Report on Speech Research *SR-25/26*, 73-86.

Leanderson, R. and B. E. F. Lindblom. (1972) Muscle activation for labial speech gestures. Acta Otolaryngol. *73*, 362-373.

Leanderson, R., A. Persson, and S. Öhman. (1971) Electromyographic studies of facial muscle activity in speech. Acta Otolaryngol. *72*, 361-369.

Lehiste, I. (1970) *Suprasegmentals*. (Cambridge, Mass.: MIT Press).

Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Amer. *35*, 1773-1781.

Linell, P. (1973) On the phonology of the Swedish vowel system. Studia Linguistica *27*, 1-2, 1-52.

Lubker, J., R. McAllister, and I. Carlson. (1974) Labial co-articulation in Swedish: A preliminary report. PILUS *23*, 13-26.

Lyttkens, I. A. and F. A. Wulff. (1885) Svenska språkets ljudlära och beteckningslära jämte en afhandling om aksent. Lund.

Malmberg, B. (1956) Distinctive features of Swedish vowels: Some instrumental and structural data. In *For Roman Jakobson*, ed. by M. Halle, H. G. Lunt, H. McLean, and C. H. van Schooneveld. (The Hague: Mouton).

McAllister, R., J. Lubker, and J. Carlson. (1974) An EMG study of some characteristics of the Swedish rounded vowels. J. Phonetics *2*, 267-278.

Noreen, A. (1903-07) *Vårt Språk I och II*. Lund.

Öhman, S. (1966) Generativa regler för det svenska verbets fonologi och prosodi. Förhandlingar vid sammankomst för att dryfta frågo rörande svenskans beskrivning III, ed. by S. Allén. Göteborg, pp. 71-87.

Port, D. K. (1971) The EMG data system. Haskins Laboratories Status Report on Speech Research *SR-25/26*, 67-72.

Port, D. K. (1973) Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research *SR-33*, 173-183.

Raphael, L. and F. Bell-Berti. (in press) The extrinsic and intrinsic tongue musculature and the feature: Tension in American English vowels. Phonetica.

Scharf, D. J. (1964) Vowel duration in whispered and in normal speech. Lang. Speech *7*, 89-97.

Tatham, M. A. A. and K. Morton. (1969) Some electromyographic data towards a model of speech production. Lang. Speech *12*, 39-53.

Teleman, U. (1969) Böjningssuffixens form i nusvenskan. Arkiv för nordisk filologi *84*, 163-208.

A Combined Cinefluorographic-Electromyographic Study of the Tongue During the Production of /s/: Preliminary Observations*

Gloria J. Borden[+] and Thomas Gay[++]
Haskins Laboratories, New Haven, Conn.

The production of the voiceless fricative /s/, especially as it occurs in two and three consonant clusters, is perhaps the most demanding tongue gesture in spoken English. It is late in normal phonological development and it is quick to deteriorate under adverse circumstances whether pathological, such as with even a mild dysarthria, or experimental, as in a temporarily induced nerve block. The production of /s/ has been examined by X-ray films, by air pressure studies, and by acoustic analysis. The study, of which this paper is the first report, is designed to investigate the organization of motor commands to the tongue muscles in the normal production of /s/ both as a single consonantal phone and as it appears in combination with other consonants. To this end, we seek to explore the interrelationships of muscle activity, tongue movement, and the resultant acoustic signal.

Our preliminary observations are based on an analysis of X-ray movies and electromyographic (EMG) recordings of our first subject. For the cinefluorography, a 16-mm cine camera recorded X-ray films at 60 frames per second. The generator delivered X-ray pulses to a 6-in image intensifier tube. Barium sulfate cream was used as a contrast medium, and several #6 BB shots were glued to the tongue tip and dorsum of which only 2 remained in place for the experiment. Details of the instrumentation may be found in Gay, Ushijima, Hirose, and Cooper (1974).

The subject read a list of utterances in which /s/ occurred in initial and final positions of a syllable and in two and three consonant clusters with plosives /sp/, /spr/, /st/, /str/, /sk/, and /skr/. The stressed syllabic nucleus was /i/, /ɑ/, or /u/, and each utterance contained a /p/ for easy identification of both X-ray movies and EMG graphs.

For the EMG recordings, hooked-wire electrodes were inserted into the following tongue muscles: the genioglossus, the superior longitudinal, the inferior

---

*This paper is based on an oral presentation at the annual convention of the American Speech and Hearing Association, Las Vegas, Nev., 5-8 November 1974.

[+] Also City College, City University of New York.

[++] Also University of Connecticut Health Center, Farmington.

[HASKINS LABORATORIES: Status Report on Speech Research SR-41 (1975)]

longitudinal, and the middle intrinsics, and, for reference, the orbicularis oris (see Hirose, 1971). After the short combined X-ray and EMG run, a longer run of EMG alone was recorded with the list of 48 utterances repeated 10 times. These two runs were analyzed separately.

The analysis of the movement data required frame-by-frame tracing of the image as projected by a Perceptoscope. The two pellets, one on the tongue tip and one approximately halfway back to the terminal sulcus on the dorsal midline, were marked on each frame and later measured as X-Y coordinates in order to graph their relative fronting and elevation. The EMG recordings were analyzed according to the Haskins Laboratories computer averaging system. For details of this system, see Kewley-Port (1973).

At first glance, there were three observations that seemed noteworthy.

(1) First of all, this subject produces /s/ with the tip of the tongue behind the lower gum ridge and the dorsum of the tongue elevated to form the constriction. Figure 1 shows the tongue tip resting behind the lower gum ridge and the dorsum bunched up. This configuration of the tongue for /s/ is consistent for this subject. The tongue tip remains fixed during /s/ but the dorsum reflects the phonetic environment of the sibilant. Although it was well-known that this alternate /s/ production occurs for many speakers, it remained to be seen what pattern of muscle activity accompanies this alternate production.

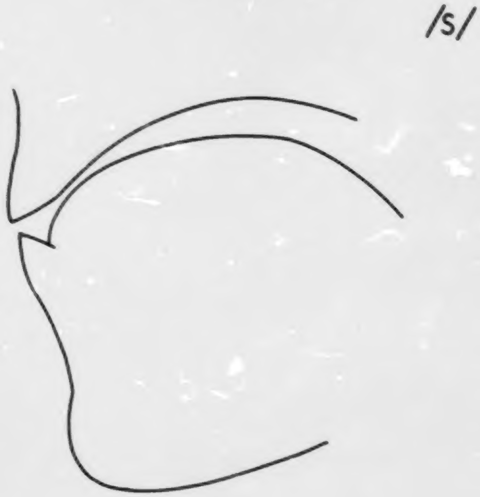


Figure 1: Alternate configuration of the tongue for sustained /s/ production. Pellets on the tip and mid-dorsum of the tongue are indicated.

For that information, we look to the EMG recordings (Figure 2) and we observe at least two muscles whose activity can be associated with the production of /s/, the inferior longitudinal and the middle intrinsics. Here the utterance is "ɑsɑpə." The horizontal boxes below the graphs indicate the actual duration of each segment of the utterance as measured from sound spectrograms. The top

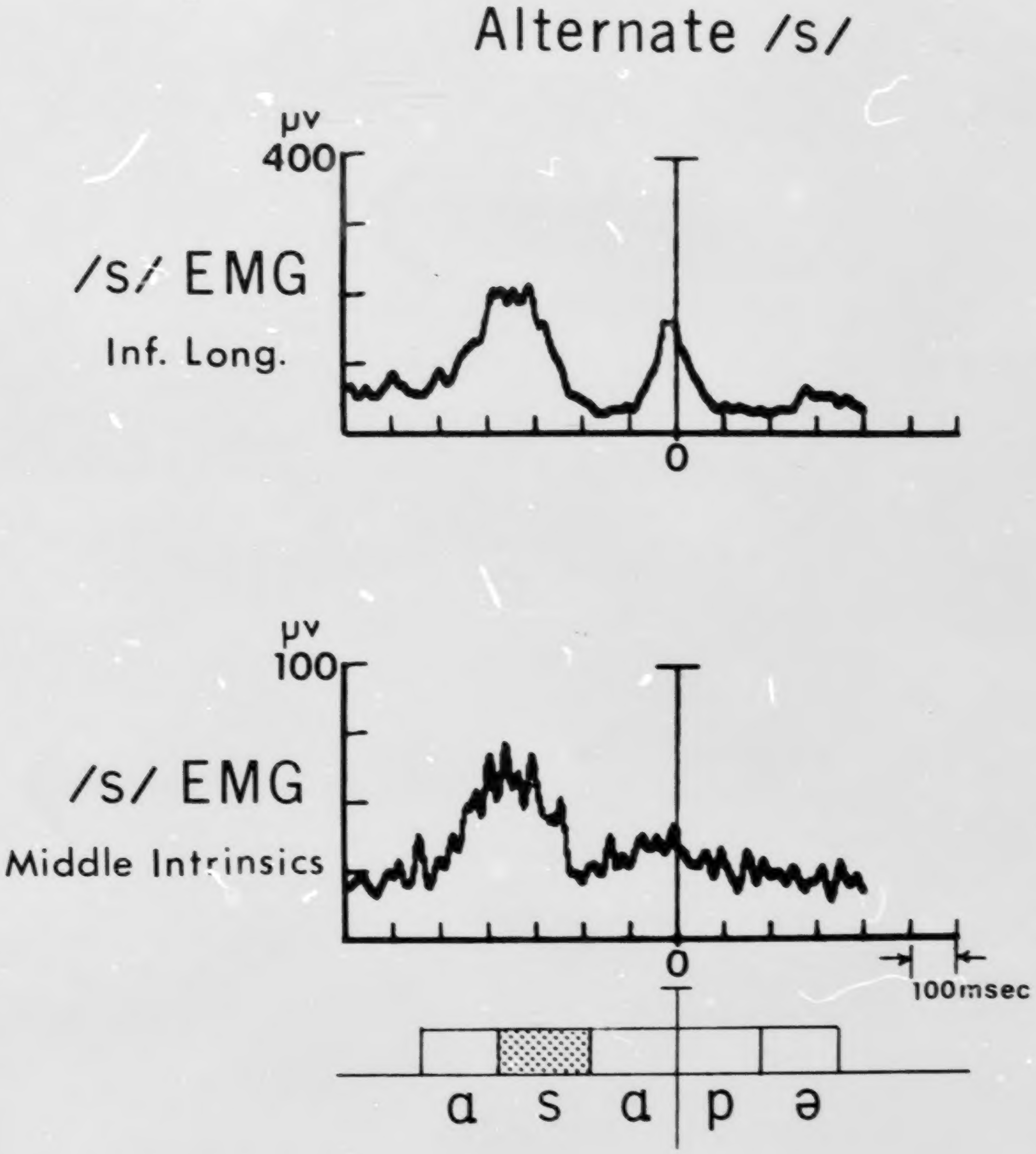


Figure 2: EMG from the inferior longitudinal muscle and in the middle intrinsic muscles of the tongue during alternate /s/ production.

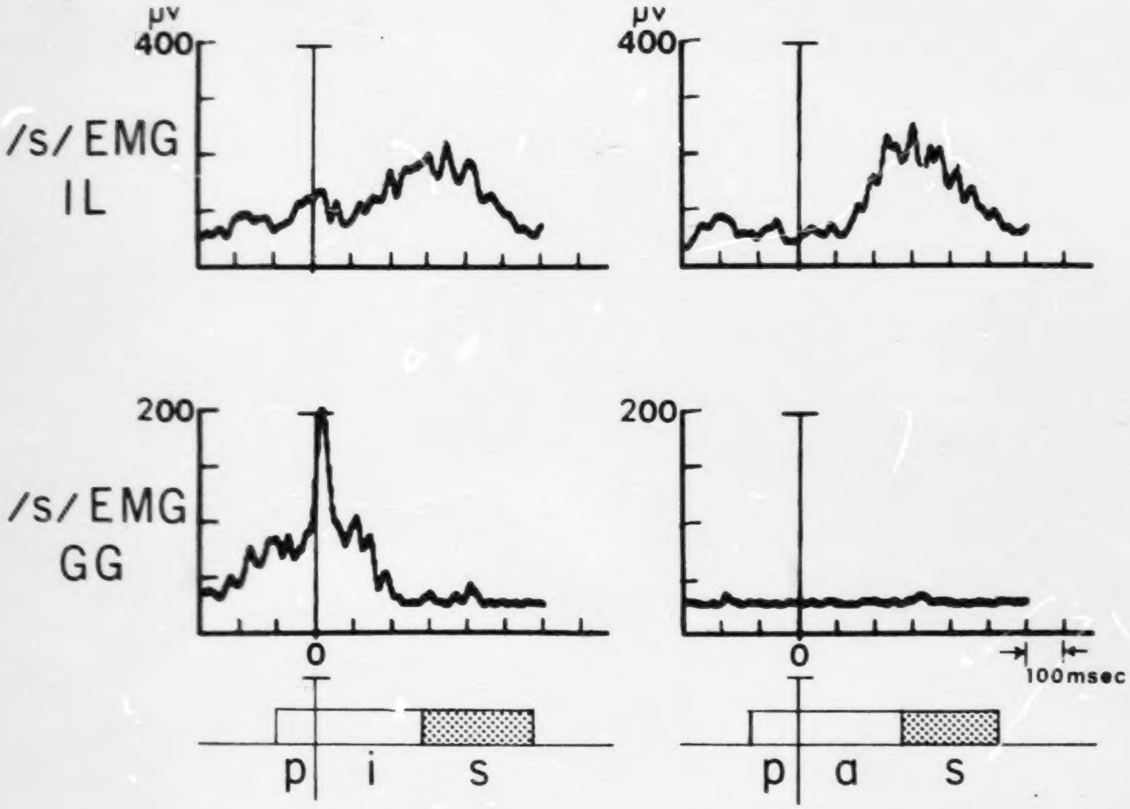


Figure 3: EMG from the inferior longitudinal muscle, which is active for /s/, and from the genioglossus muscle, which peaks for /i/.

graph shows the buildup of electrical potential from the contraction of the inferior longitudinal muscle, starting in this case about 100 msec before the acoustic event of the fricative during which it peaks. The inferior longitudinal muscle courses along the inferior aspect of the tongue and its contraction curves the tongue tip downward. It is apparent, then, that this low tongue tip /s/ is not passive, the result of being left behind when the dorsum elevates, but is the result of an active gesture of apical depression to facilitate the bunching of the tongue. The second graph is also indicative of a pattern that is consistent with the occurrence of /s/, a pattern of activity of the middle intrinsic muscles of the tongue. For this subject, then, the middle intrinsic muscles and the inferior longitudinal muscles were consistent in their contraction for the production of the alternate /s/.

Again, in Figure 3 the active apical depression by the activity of the inferior longitudinal can be seen in the top two graphs for the /s/ in syllable-final position, in /pis/ and /pɑs/. The lower pair of graphs represents the level of activity as recorded from the genioglossus muscle. It is apparent in these utterances that the genioglossus muscle is active for /i/ but not for /ɑ/, a finding that is common and that relates to our second observation concerning /s/ clusters.

(2) As we see in Figure 4, the target shape for /i/ produced by this subject is remarkably stable whether it is preceded by /sp/, /st/, or /sk/.

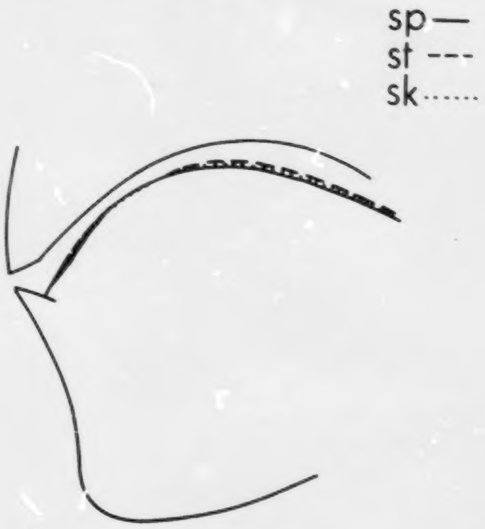


Figure 4: Tongue configuration for /i/ after /sp/, /st/, and /sk/.

The EMG signals (Figure 5), in contrast to the movement data, show that the contraction of the genioglossus for /i/ varies in its relative level of activity depending on whether the /i/ follows /sp/, /st/, or /sk/. The column of graphs on the left-hand side are of /spipə/, /stipə/, and /skipə/, on the right-hand side, /spripə/, /stripə/, and /skripə/. Notice how the activity of the

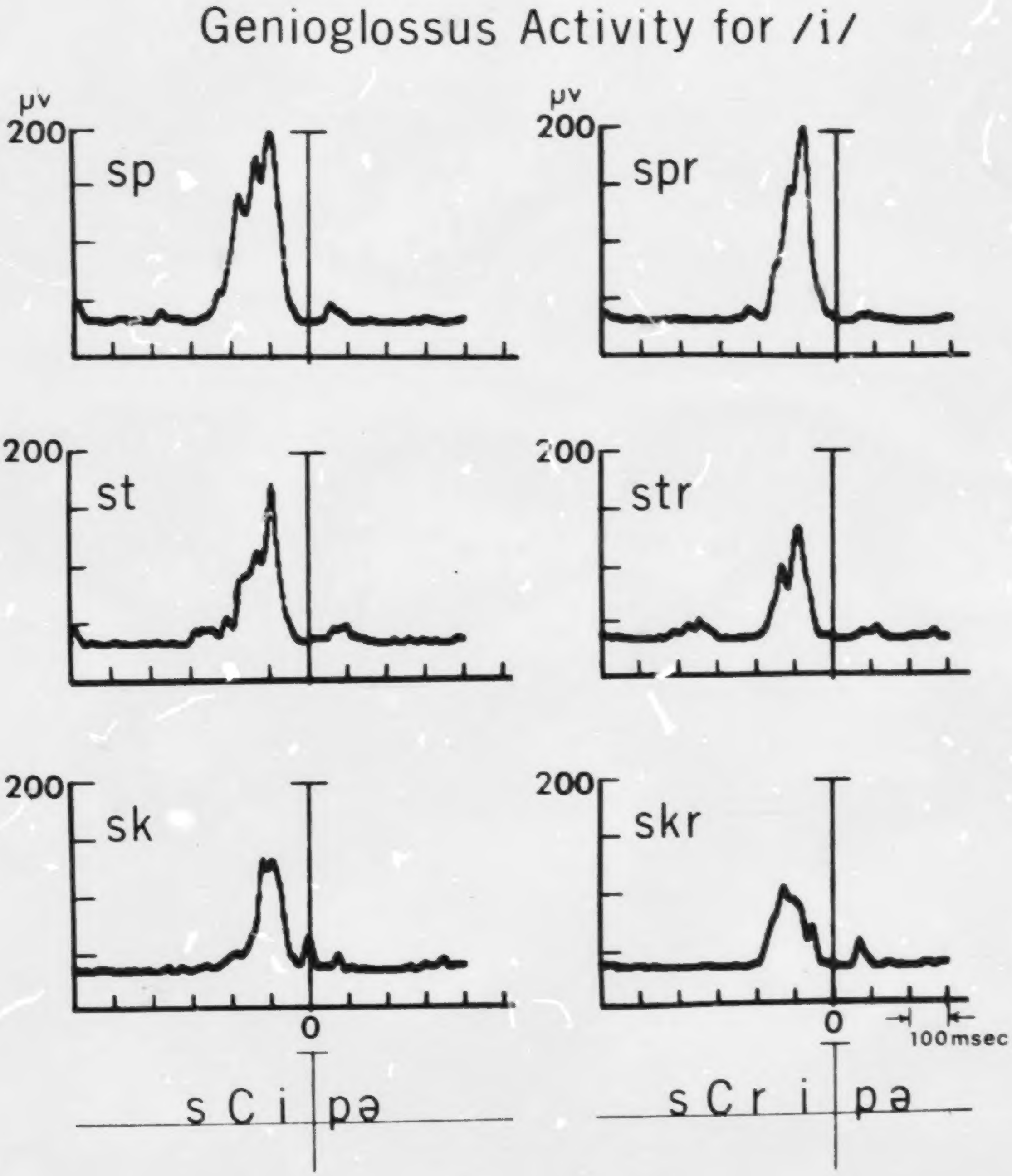


Figure 5: EMG from the genioglossus muscle for /i/ in several phonetic contexts.

genioglossus muscle is diminished after /sk/ and /skr/. It may be that the genioglossus effectively lifts the tongue for /i/ after /sp/ and /st/ but is aided by the mylohyoid or styloglossus muscles for tongue lifting for /k/; therefore, fewer demands are put on the genioglossus in this context.

So we observe that although the movement data indicate that the target shape is the same for /i/ whether after /sp/, /st/, or /sk/, the muscles used to obtain this target vary according to the preceding consonants. This is an example of how there are not always invariant motor commands for each phoneme, but rather each command may be interlaced with other muscle commands for segments as large as /skri/ in this case, a four-phone syllable.

(3) The last observation to report from these data concerns duration. Another way to approach the question of how the motor commands are organized for speech is by looking at durational differences. Schwartz (1970), Klatt (1974), and others have reported that acoustically an /s/ is shorter before /p/ than it is before /t/ or /k/ in a consonant cluster, so in /sp/, the /s/ is shorter than in /st/ or /sk/. Lining up the acoustic signal with the movement and EMG data should tell us whether this durational difference is true on the movement level or on the motor command level. Is there an invisible inaudible /s/ in /sp/ having a duration similar to that in /st/ and /sk/ but simply occluded by the /p/?

We've just started to look at these relationships but the durational differences seem to hold up in the movement data. The tongue position for /s/ before /p/ is held only to the /p/ closure when it starts to move toward the vowel target. Therefore, the target shape for /s/ is shorter in /sp/ than for /st/ and /sk/.

Since the inferior longitudinal muscle was so consistent in its activity for /s/ in this subject, we looked there for the durational difference on the motor command level.

In Figure 6 one can see that the inferior longitudinal is active for the final /s/ clusters in the left column of graphs. There is a tendency for the activity to fall off more sharply for the /s/ in /sp/ than in /st/ or /sk/ where it is maintained longer. The arrows on the figure point to the slope differences. The durational difference, then, also operates on the EMG level. The same thing happens for initial clusters as seen in the right column of graphs. The falloff of activity is steeper for /sp/ than for the other clusters.

With labial closure, the tongue is free to move on to the vowel target, but for /st/ or /sk/, the tongue is involved throughout the cluster. In other words, the tongue is free to coarticulate with the lips for /sp/ in anticipation of the vowel but not for /st/ or /sk/, since the tongue is delayed with its involvement in the /t/ and /k/ gestures.

In conclusion, we find that by simultaneously viewing movement data and muscle activity data, we can observe evidence of different muscles contracting for the same target shape and phone depending upon which phones precede it. We can observe evidence of the freedom to coarticulate within a syllable as the tongue is free to move toward /ɑ/ during the labial closure in /sp/. Finally, we have EMG evidence as well as movement evidence that there is more than one way to produce a sound that is acoustically acceptable and well within the phoneme boundaries of /s/, an alternate /s/.

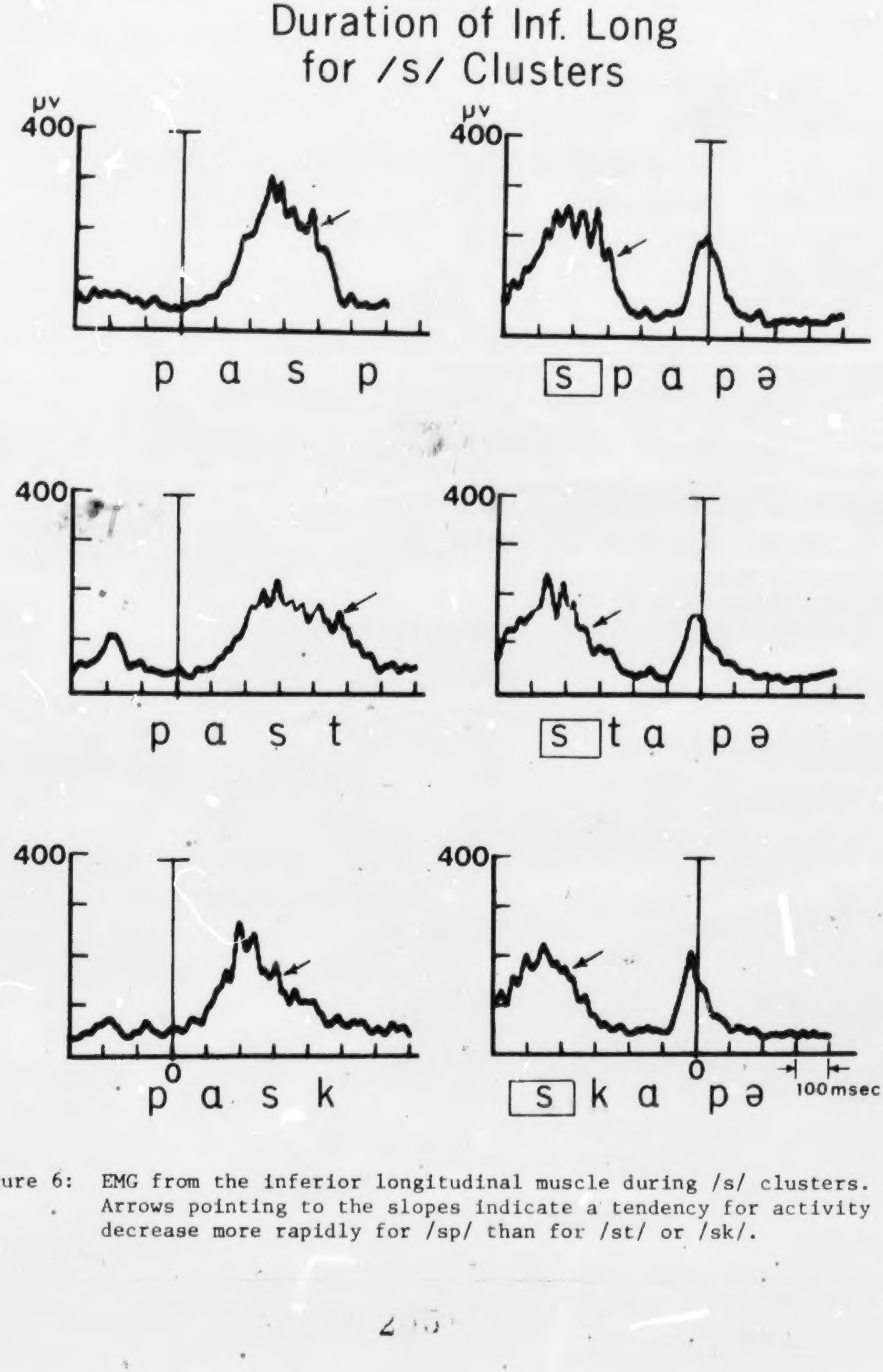


Figure 6: EMG from the inferior longitudinal muscle during /s/ clusters. Arrows pointing to the slopes indicate a tendency for activity to decrease more rapidly for /sp/ than for /st/ or /sk/.

We have filmed two more subjects who have tongue-tip high /s/ and are planning another experiment involving simultaneous EMG and cine X-rays of a fourth subject, also with a high apical /s/. It will be interesting to compare those data when they are processed with this subject to get an idea of subject variation in the motor organization of /s/ and /s/ clusters.

REFERENCES

Gay, T., T. Ushijima, H. Hirose, and F. S. Cooper. (1974) Effect of speaking rate on labial consonant-vowel articulation. J. Phonetics 2, 47-63.

Hirose, H. (1971) Electromyography of the articulatory muscles: Current instrumentation and technique. Haskins Laboratories Status Report on Speech Research SR-25/26, 73-86.

Kewley-Port, D. (1973) Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research SR-33, 173-184.

Klatt, D. (1974) The duration of [s] in English words. J. Speech Hearing Res. 17, 51-63.

Schwartz, M. F. (1970) Duration of /s/ in /s/-plosive blends. J. Acoust. Soc. Amer. 47, 1143(Letter).

Velar Movement and Its Motor Command*

T. Ushijima[+] and H. Hirose[+]
Haskins Laboratories, New Haven, Conn.

In order to investigate the relationship between movement of the velum and its motor command during speech, electromyographic (EMG) recordings of the levator palatini muscle and direct viewing of the velum were performed separately on one Japanese subject. The same utterance types were used in both experiments.

Electromyographic signals were computer-processed and will be shown in the form of an averaged and smoothed pattern for each utterance type (Kewley-Port, 1973). Velar movement was filmed at the rate of 50 frames per second through a fiberscope inserted in the nasal cavity (Ushijima and Sawashima, 1972).

In this report we would like to point out four important findings obtained in this study.

The lower part of Figure 1 shows the time course of velar height for the utterance /seesee/ followed by a carrier word /desu/.[1] Note the difference in height between the consonants and the vowels. Similarly, the level of EMG activity associated with the interconsonantal vowel /e/ is much lower than for /s/ in the upper figure (thick line). This difference is quite consistent for all the samples. This implies that for this particular subject the different levels of EMG activity between /s/ and /e/ seem to be realized in the form of small differences in velar height. In other words, there seem to be quantitatively different neural commands for the movements of the velum for consonant and vowel production, although both segments are generally regarded as [-] nasality. It seems reasonable to say that the velum is not controlled by a simple binary on-off mechanism.

The next point is related to the differences among four nonnasal consonants. Figure 2 shows comparisons of peak EMG amplitude for the consonants /t/, /s/, /d/, and /z/ in each utterance type. They are classified and pooled into seven groups according to their phonetic environment. It should be

---

*Paper presented at the 1974 annual convention of the American Speech and Hearing Association, Las Vegas, Nev.

[+] Also University of Tokyo, Japan.

[1] In this figure "'", "#", and "N" represent, respectively, a syllable boundary, a word boundary, and a syllable-final nasal.

[HASKINS LABORATORIES: Status Report on Speech Research SR-41 (1975)]

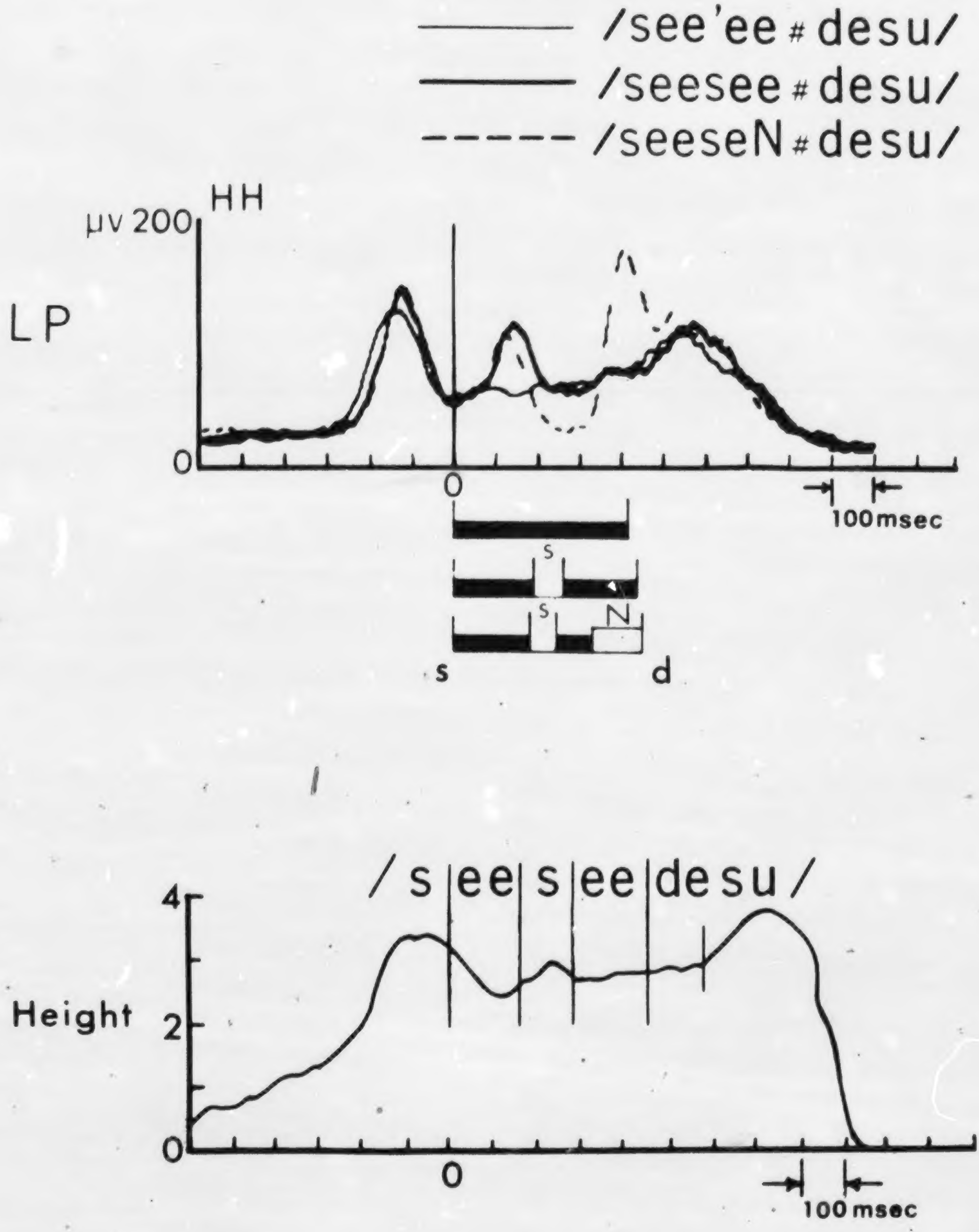
/see'ee # desu/
/seesee # desu/
/seeseN # desu/
LP
HH
µv 200
0
0
100msec
s
S
N
s
d
/ s | ee | s | ee | de su /
Height
4
2
0
0
100 msec

FIGURE 1

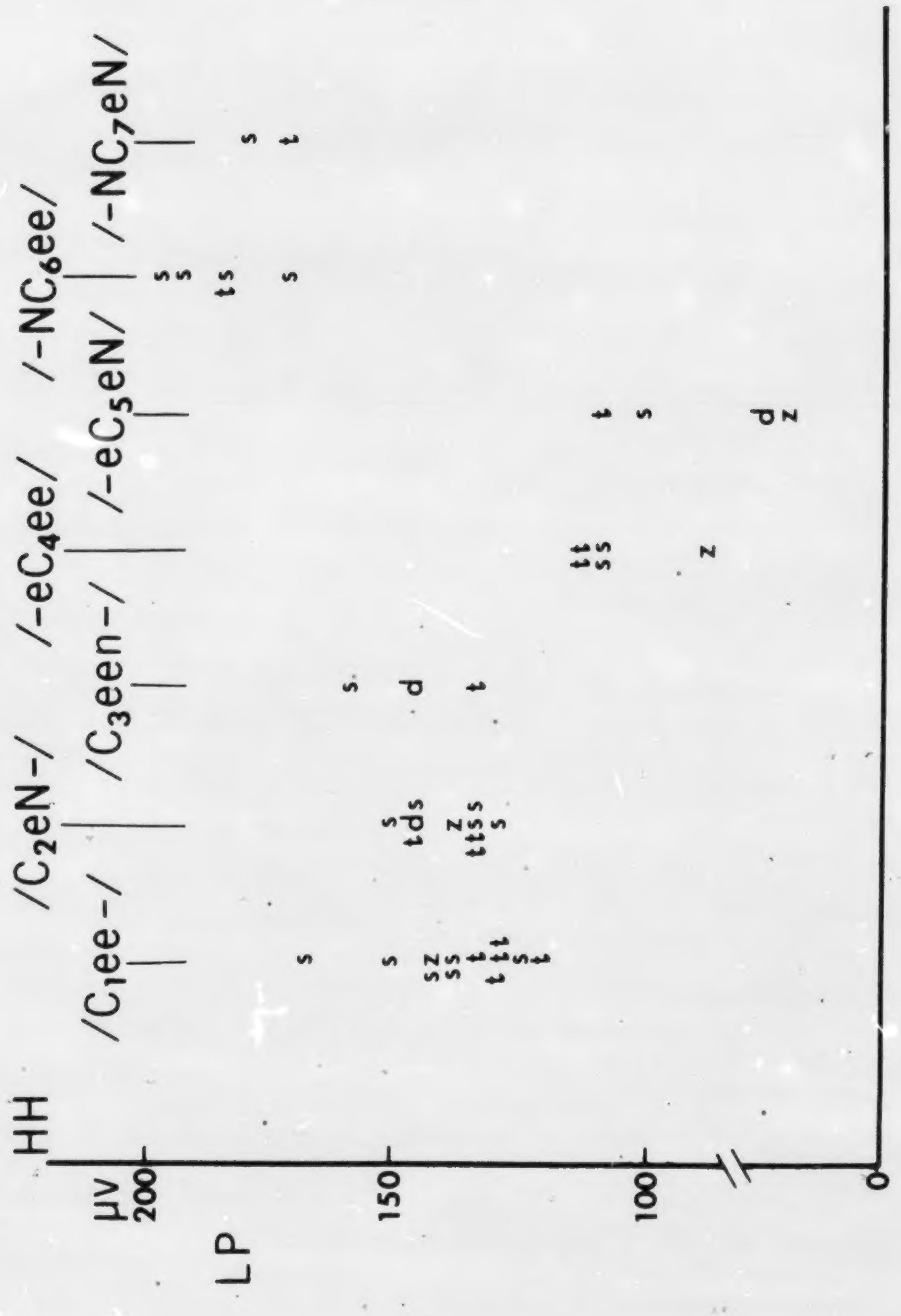

HH
LP
μV
200
150
100
0
/C₁ee–/
/C₂eN–/
/C₃een–/
/–eC₄ee/
/–eC₅eN/
/–NC₆ee/
/–NC₇eN/
s
z
t
d

FIGURE 2

remarked here that there is variation in the peak values for the utterances in each group. There is no clear systematic segmental difference between voiced and voiceless consonants or between stop and fricative consonants.[2] Instead, it is notable that the activity is greatest for the consonants after a syllable-final nasal /N/ (Groups 6 and 7), and least for the consonants in intervocalic position (Groups 4 and 5). It seems, then, that the activity level of the muscle for a given nonnasal phoneme varies according to its phonetic environment.

The same comparisons may also be obtained for velar height (Figure 3). The environment of the consonant seems to be the most dominant factor in determining velar height. The discrepancy between EMG measures and velar height measures for consonants following nasals is due to the fact that EMG activity is related to the distance the palate must move, rather than the absolute height it reaches. At any rate, our data seem to show that the activity level of the muscle and velar height for a given nonnasal phoneme vary according to its phonetic environment.

The third point concerns differences between syllable-initial and syllable-final nasals. In Japanese, there are two syllable-initial nasals, one labial, one alveolar. However, a syllable-final nasal (the uppercase /N/) has some special features. Its inherent phonetic values, nasality and voicing, are prerequisite, but the specification of its place of articulation varies as a function of the following phoneme (Fujimura, 1972). For example, labial, alveolar, and velar articulations are possible.[3]

The upper part of Figure 4 shows the EMG curves, while the lower part shows velar height curves for the contrastive pair. Our earlier data, obtained from velar movement analysis using other subjects, implied an inherent difference between the two nasals with greater velar suppression for the syllable-final nasal (Ushijima and Sawashima, 1972). We also reported some EMG evidence supporting that result, and commented that the duration of nasal segments and speaking rate may be important factors for determining velar height (Ushijima and Hirose, 1974).

In Figure 5 we have plotted velar height and the duration of the acoustic segment for each nasal occurrence in the fiberoptic run in this study. The

---

[2] In this figure the voiced consonants fail to show a constantly higher elevation of the velum than their voiceless counterparts. One reason for this result seems related to the fact that the levator activity, or velar height, is not the only indication of pharyngeal cavity enlargement. The strategy for pharyngeal enlargement to maintain voicing probably differs subject by subject as Bell-Berti and Hirose (1972) first reported. This particular subject seems to use strategies other than the velar height change for voicing.

[3] Labial closure occurs for a /N/ if it is followed by a labial consonant. Dental or alveolar closure occurs before /t/ or /d/. More posterior place of closure is seen if the phoneme is followed by /s/, /z/, /k/, or /g/. There is less vocal-tract stricture if the phoneme precedes a vowel. In such a case the syllable-final nasal becomes phonetically equivalent to a nasalized vowel.

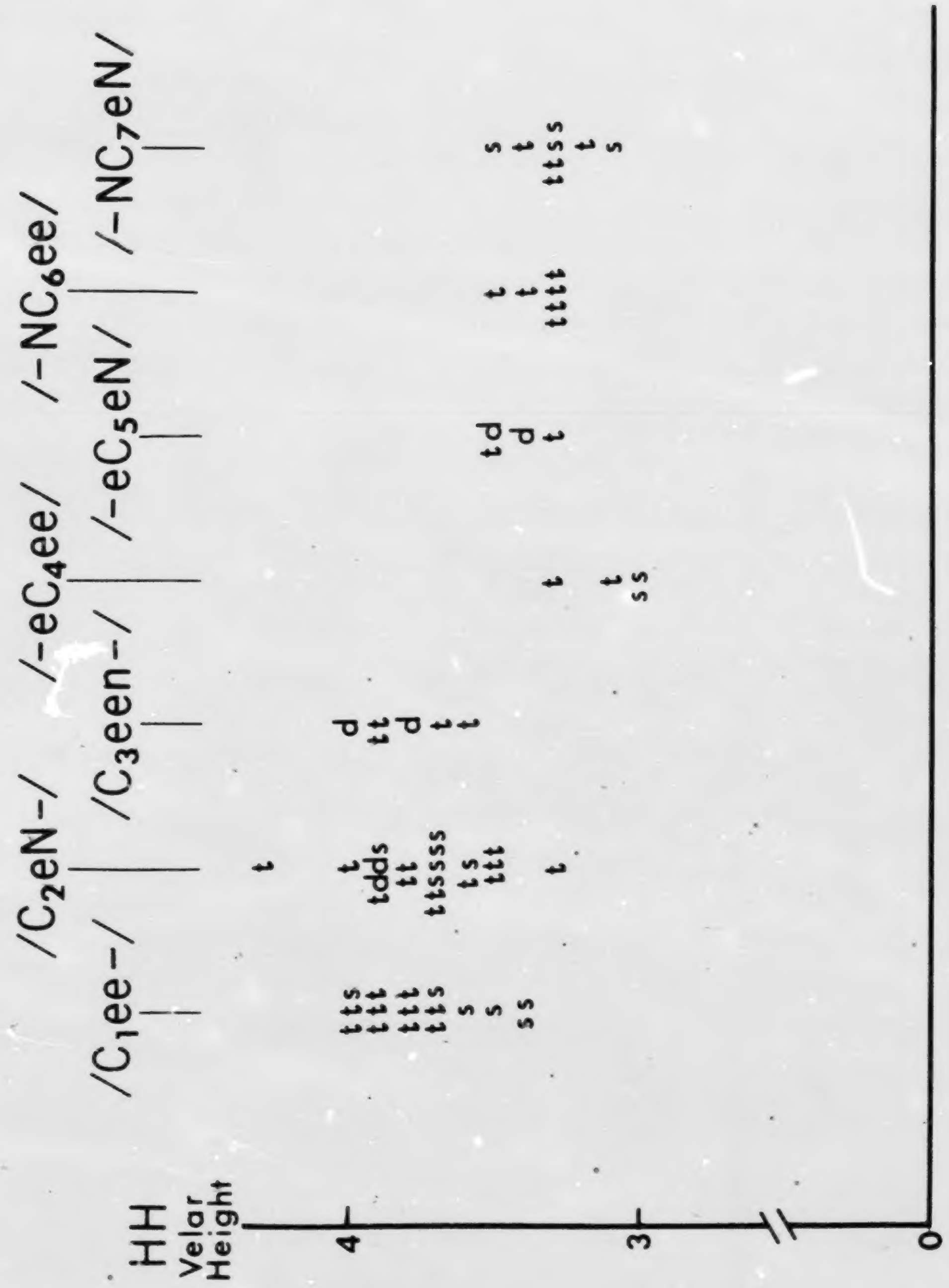
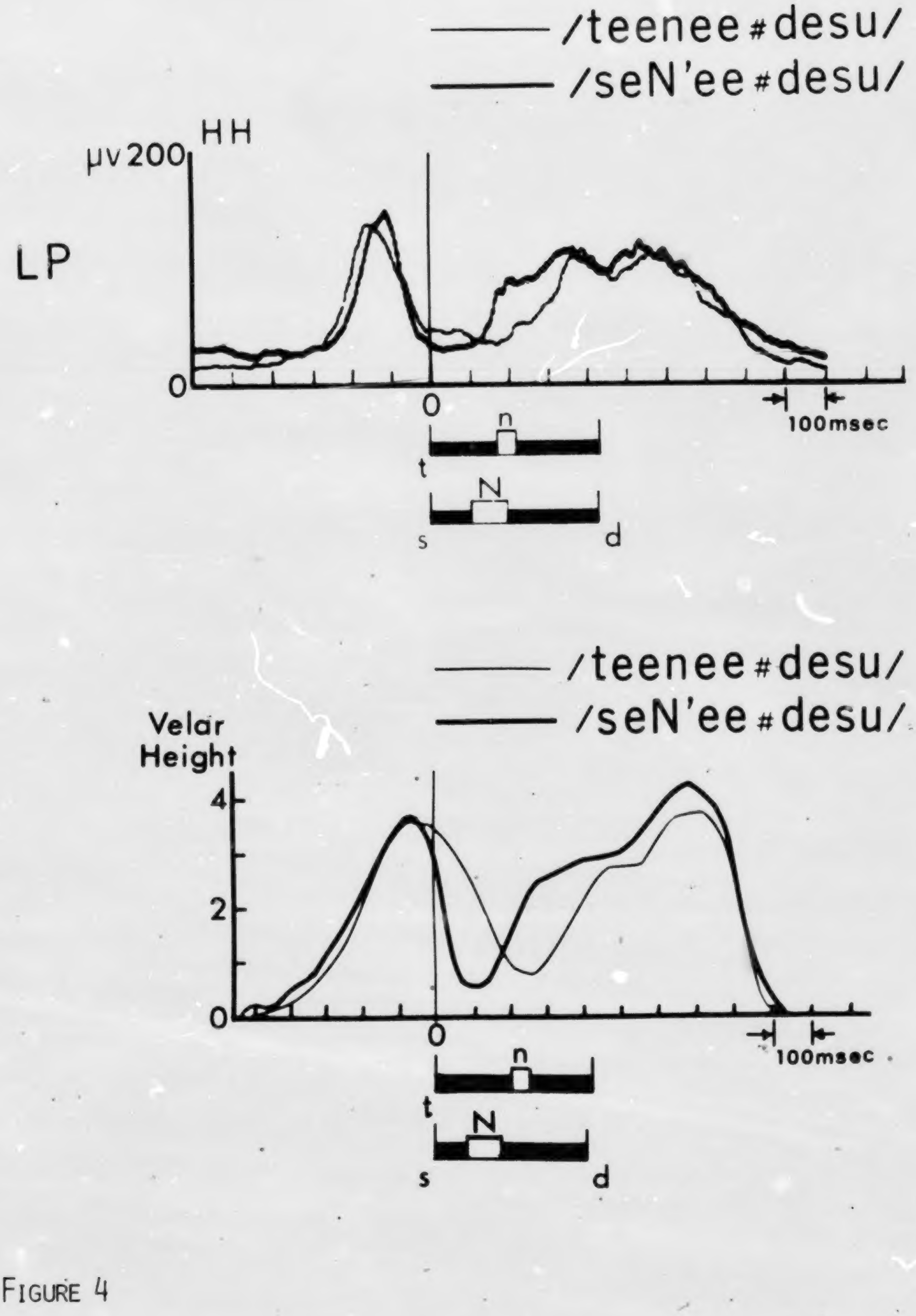


FIGURE 3



/teenee#desu/
/seN'ee#desu/
HH
μv200
LP
0
0
100msec
n
t
N
s
d
/teenee#desu/
/seN'ee#desu/
Velar Height
4
2
0
0
100msec
n
t
N
s
d

FIGURE 4

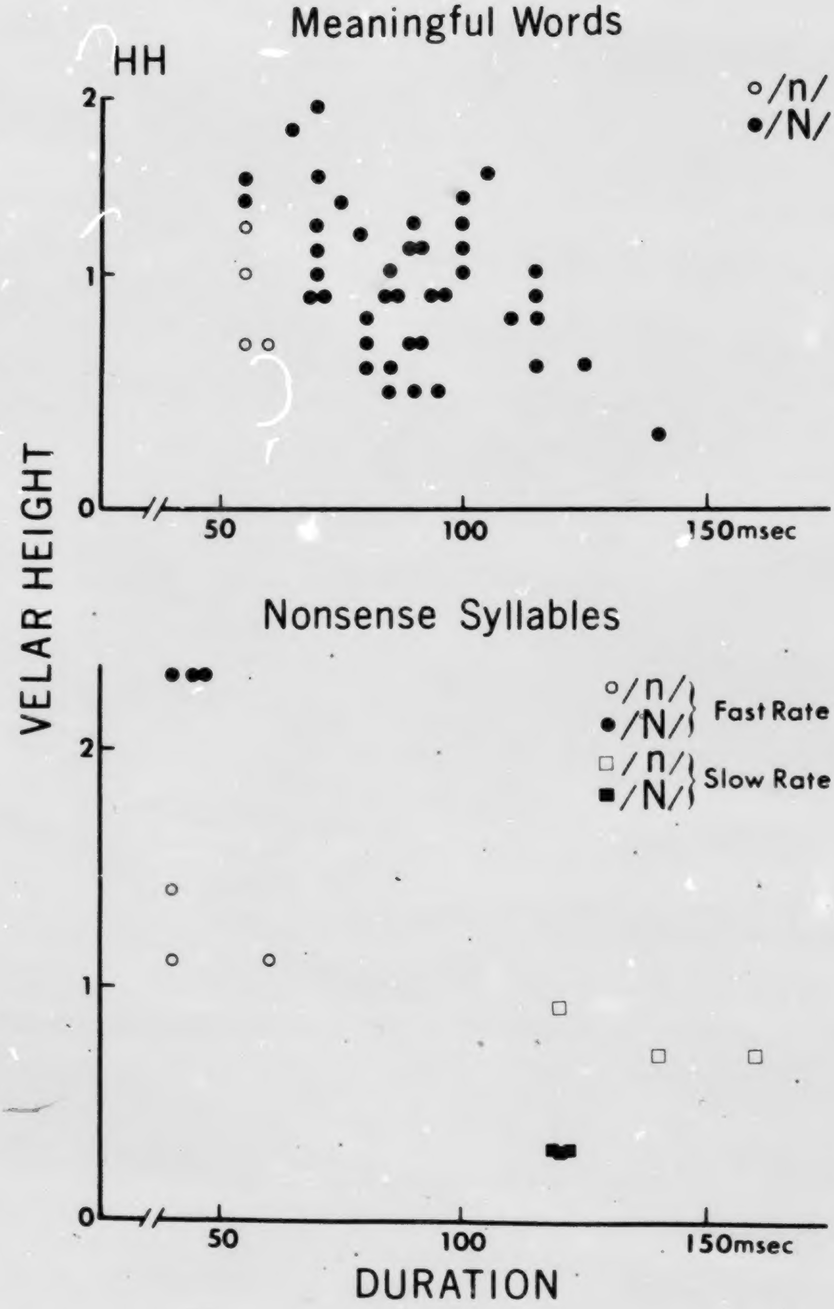

Meaningful Words
HH
○/n/
●/N/
2
1
0
50
100
150msec
VELAR HEIGHT
Nonsense Syllables
○/n/ ●/N/ Fast Rate
□/n/ ■/N/ Slow Rate
2
1
0
50
100
150msec
DURATION

FIGURE 5

upper part of the figure was made from samples of meaningful words, spoken at conversational rate. The lower part of the figure shows the effect of speaking rate on velar height and duration, since both the initial and final nasals were repeated in a string of nonsense syllables at both fast and slow speeds. It is interesting to note that velar height for the final nasal /N/ (filled circles and squares) tends to vary with both duration and speaking rate, while the initial nasal /n/ (open circles or squares) seems to be more independent of these two factors.

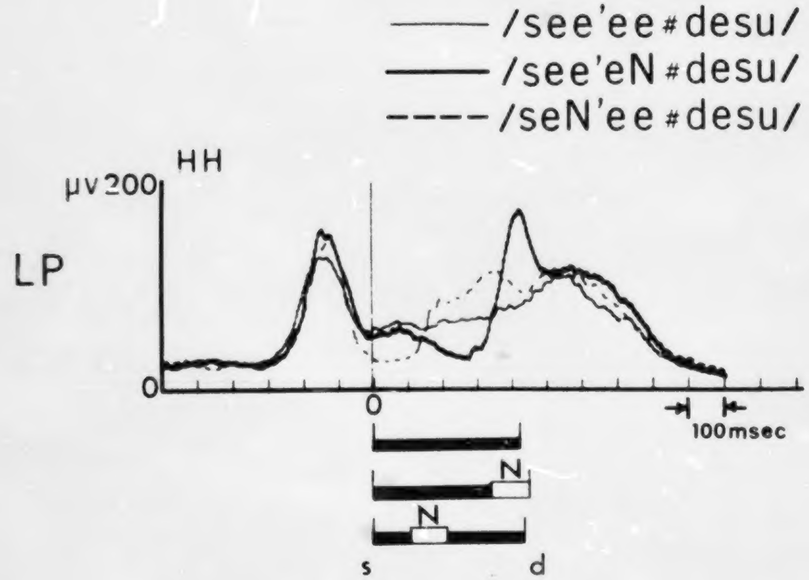
A possible explanation for this might be the following: articulatory accuracy is required for the production of the initial nasal /n/, that is, complete contact of the tongue tip to the alveolar ridge with simultaneous lowering of the velum. On the other hand, for the production of the final nasal /N/, the place of articulation tends to be less constant as the duration becomes short, which might cause less lowering of the velum. Of course, such a hypothesis should be clarified by further studies using other methods such as spectrography, cineradiography, or palatography.

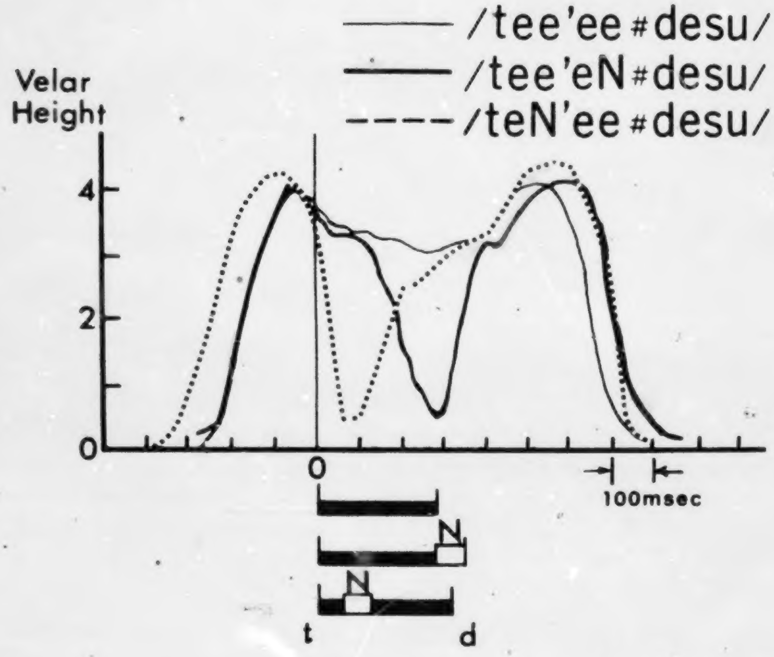
The final point we would like to make is about nasal coarticulation of the velum. The analysis of velar movement in this study provides us with results that support our previous EMG data (Ushijima and Hirose, 1974).

In Figure 6 we compare three utterance types. The thin line represents an entirely oral sequence /see'ee/. The thick line represents a sequence /see'eN/ with a syllable-final nasal in final position in the test word. The dashed line represents a sequence with a syllable-final nasal in medial position, /seN'ee/. The dashed line in the upper figure shows that immediately after the peak for /s/ there is suppression of EMG activity for the vowel and following nasal. In contrast, in the utterance with the sequence-final nasal (the thick line), activity for the vowel after /s/ has the same level as the vowel in the utterance without the nasal (the thin line). Looking at the thick line, we see that the activity begins to fall about 100 msec after the lineup. The lower figure also indicates the clearly delayed onset of the velar lowering for the sequence-final nasal /N/ (the thick line) compared with the sequence-medial nasal /N/ (the dashed line). This phenomenon might be interpreted as indicating that there is a restriction on anticipatory lowering of the velum. Moll and Daniloff (1971) proposed a hypothesis of "unspecified" velar position for English vowels. According to them, velar lowering for a terminal nasal should start at the beginning of a preceding vowel string.

In this sense, our data do not appear to support their hypothesis. It seems reasonable to consider that anticipatory coarticulation may not always extend beyond a syllable boundary.

If we again examine the dashed line in the upper half of Figure 6, we see that EMG activity for the vowel after a syllable-final nasal /N/ does not reveal any carry-over suppression from the preceding nasal. Rather, it shows an increase over the EMG level necessary for the vowel sounds of the completely oral sequence (the thin line). At the level of the neural command to the velum, then, there is no carry-over effect in the vowel segment after the syllable-final nasal /N/. In this case, then, the carry-over effect does not seem to extend beyond the syllable boundary of the test word. This tendency is also visible in the lower figure showing velar height. One possbble explanation for


/see'ee#desu/
/see'eN#desu/
/seN'ee#desu/
HH
μv 200
LP
0
0
100msec
N
N
s
d


/tee'ee#desu/
/tee'eN#desu/
/teN'ee#desu/
Velar Height
4
2
0
0
100msec
N
N
t
d

this phenomenon is that the elongation of the vowel segments after a syllable-final nasal /N/ may have to be oralized to prevent listener confusion.[4] On the other hand, we observed a clear carry-over effect for the vowel segment following a syllable-initial nasal /n/, which is not shown in Figure 6.

Our observations have been based entirely on Japanese materials, but we assume some of the results of this study can be generalized to other languages.

REFERENCES

Bell-Berti, F. and H. Hirose. (1972) Stop consonant voicing and pharyngeal cavity size. Haskins Laboratories Status Report on Speech Research SR-31/32, 207-211. [Later published as Bell-Berti, F. (1975) Control of pharyngeal cavity size for English voiced and voiceless stops. J. Acoust. Soc. Amer. 57, 456-461.]

Fujimura, O. (1972) Fundamentals of speech science. In Onseikagaku (Speech Science), ed. by J. Oizumi and O. Fujimura. (Tokyo: Tokyo University Press), pp. 3-91.

Kewley-Port, D. (1973) Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research SR-33, 173-183.

Moll, K. L. and R. G. Daniloff. (1971) Investigation of the timing of the velar movements during speech. J. Acoust. Soc. Amer. 50, 678-684.

Ushijima, T. and H. Hirose. (1974) Electromyographic study of the velum during speech. Haskins Laboratories Status Report on Speech Research SR-37/38, 79-97.

Ushijima, T. and M. Sawashima. (1972) Fiberscopic observation of velar movements during speech. Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo) 6, 25-38.

---

[4] One example of possible listener confusion:

/KaN'oo/ (to enjoy seeing the cherry blossoms)

vs.

/KaNnoo/ (full payment of a tax).

The Stuttering Larynx: An EMG, Fiberoptic Study of Laryngeal Activity Accompanying the Moment of Stuttering

Frances Freeman* and Tatsujiro Ushijima+
Haskins Laboratories, New Haven, Conn.

Over a century ago Arnott (1828) wrote, "the most common cause of stuttering is in the glottis." Other writers, including Müller (1857), Hunt (1861), and Kenyon (1943), proposed models of the stuttering block incorporating a primary laryngeal component. The present research sought to test this century-old hypothesis through two methodologies. The first approach utilized a fiberoptic endoscope for direct observation of the glottis, while the second utilized multichannel electromyography to investigate the motor commands that resulted in the observed laryngeal dysfunction.

Comments on the fiberoptic studies will be brief for two reasons: first, because a film of work is currently available;[1] and second because of overlap with recent work of Contour, Brewer, and McCall (1974).

When Chevrie-Muller (1963) utilized the glottalgraph to study 27 stutterers, she reported: (1) arhythmic vocal-fold vibrations, (2) unpredictable glottal openings, and (3) partial or complete absence of voicing during rapid glottal activity. Fujita (1966) took anterior-posterior X-rays of a Japanese stutterer and reported: (1) irregular or inconsistent opening and closing of the pharyngo-laryngeal cavity and (2) asymmetric tight closure of the glottis, which extended upward and included closure of the pharyngeal cavity. Our own fiberoptic observations found: (1) irregular, unpredictable glottal openings and (2) very tight closure of the laryngeal aperture.

Figure 1 shows individual frames from the motion picture, illustrating the sequence typical of the tight laryngeal closure in some moments of stuttering. Each frame is shown in its original form, and in a tracing of the tissue outline. Frame 1 shows the true folds in phonatory position. In frame 2 the ventricular folds can be seen as they are adducted and partially occlude the

---

*Also City University of New York and Adelphi University, Garden City, N. Y.

+Also University of Tokyo, Japan.

[1] This film (Kamiyama, Hirose, Ushijima, and Niimi, 1965) was included, with subtitles, in the American Speech and Hearing Association 1974 film theater offerings. Inquiries should be addressed to Dr. Seiji Niimi, Haskins Laboratories, New Haven, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-41 (1975)]

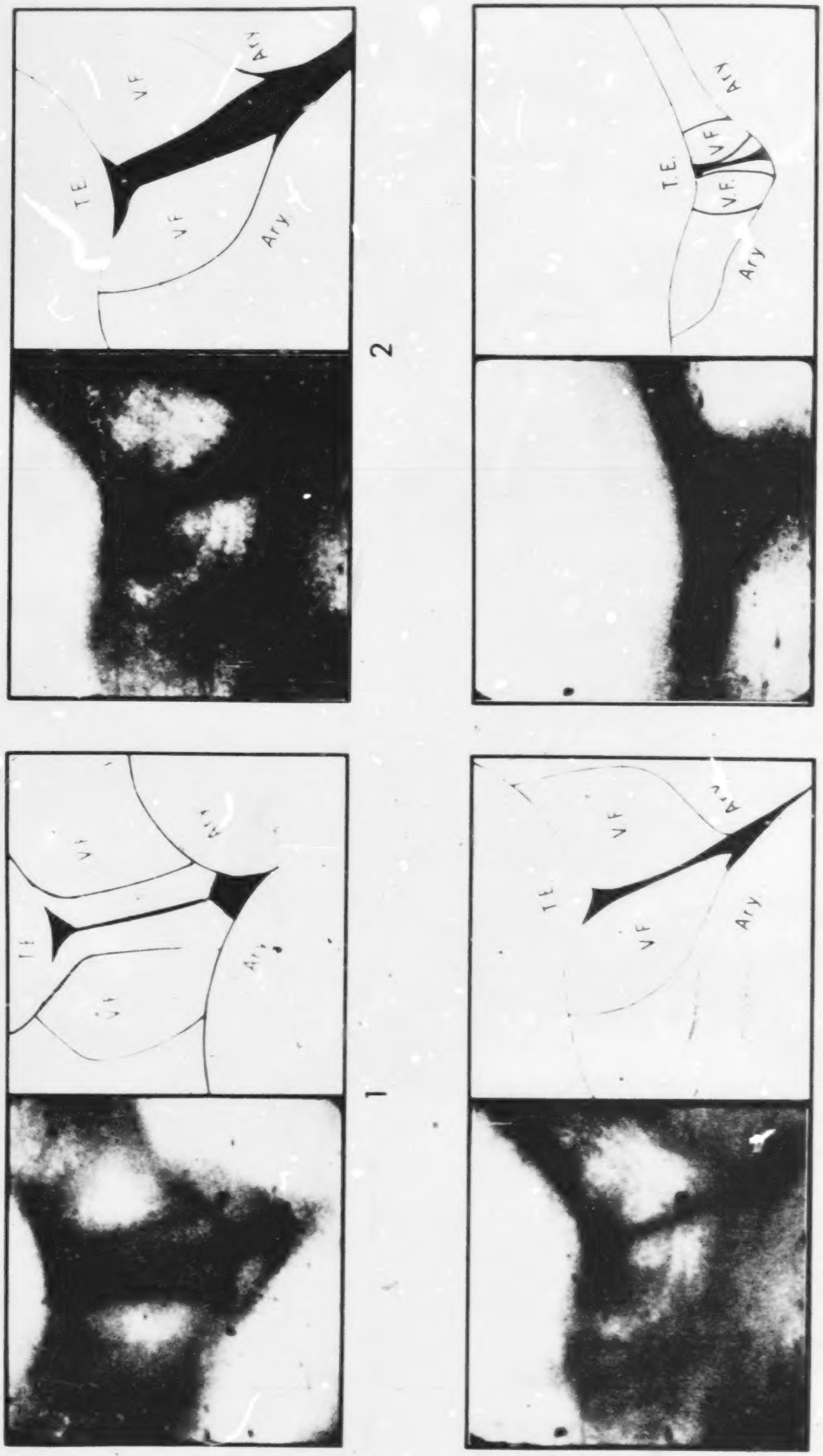

1
T.E.
VF
VF
Ary
Ary
2
T.E.
VF
VF
Ary
Ary
3
T.E.
VF
VF
Ary
Ary
4
T.E.
V.F.
V.F.
Ary
Ary

FIGURE 1

2 3

glottis. In frame 3 the adduction of the ventricular folds is almost complete. Frame 4 shows anterior-posterior closure at the level of the arytenoids and the tuberculum of the epiglottis. Even with this tight closure of the larynx, the subject is still attempting to phonate through the stricture, as the sound track accompanying the film shows. Other fiberoptic studies show that blocking also includes depression of the epiglottis.

The electromyographic (EMG) techniques used in the second phase of our work have been developed in a series of experiments investigating the normal laryngeal muscle activity in phonation and speech (Faaborg-Anderson, 1957; Hirano and Ohala, 1969; Hirano, Ohala, and Vennard, 1970; Hirose, 1971; Shipp and McGlone, 1971; Gay, Strome, Hirose, and Sawashima, 1972; Hirose and Gay, 1972, 1973). Experimental procedures are described in Hirose (1971) and data collection and processing are discussed in Port (1971, 1973, 1974).

Data were collected on three subjects. Attempts were made to record simultaneously from all five intrinsic laryngeal muscles and from three upper-tract articulatory muscles. In each case, acceptable recordings were obtained from four of the five intrinsic laryngeals (though the set varies from subject to subject) and three upper-tract articulators. Repeat recordings were made in a second session with one subject.

The first stage of data processing yielded oscillographic tracings of the muscle action potentials and the acoustic signal. Inspection of these "raw" EMG tracings yielded findings relevant to the "Wingate hypothesis." Wingate (1969, 1970) reevaluated the speaking conditions under which most stutterers are fluent. These conditions included whispering, choral speaking, speaking in rhythm, or speaking under delayed auditory feedback (DAF), or auditory masking. He hypothesized that, "in these circumstances which improve fluency, the stutterer is induced, in one way or another, to do something with his voice that he does not ordinarily do" (Wingate, 1969:684-685).

Each subject read the same material under three or more of these conditions: white noise, DAF, rhythm, choral speaking, and whispering. These conditions had the anticipated effect of reducing stuttering in the three subjects. The following data samples allow a comparison of EMG recordings taken during a stuttered reading of a sentence with those taken during the reading of the same sentence under a fluency-evoking condition.

Figure 2 shows three laryngeal muscle recordings for subject PN reading the phrase "and the origin of all false science and imposture is in the desire to accept false causes rather than none." The upper graph shows a stuttered reading and the lower shows a fluent reading under white noise. The overall activity levels are higher for the stuttered reading and lower for the fluent condition.

Figure 3 shows recordings for subject GG from the same three muscles for the same sentence. Here the fluency-inducing condition is rhythm reading. Again, the fluent condition shows activity levels that are lower than those recorded in the disfluent reading.

Figure 4 shows data from subject DM for posterior cricoarytenoid (PCA), vocalis (VOC), and lateral cricoarytenoid (LCA). The effects of choral reading on the activity levels of these muscles are particularly dramatic.

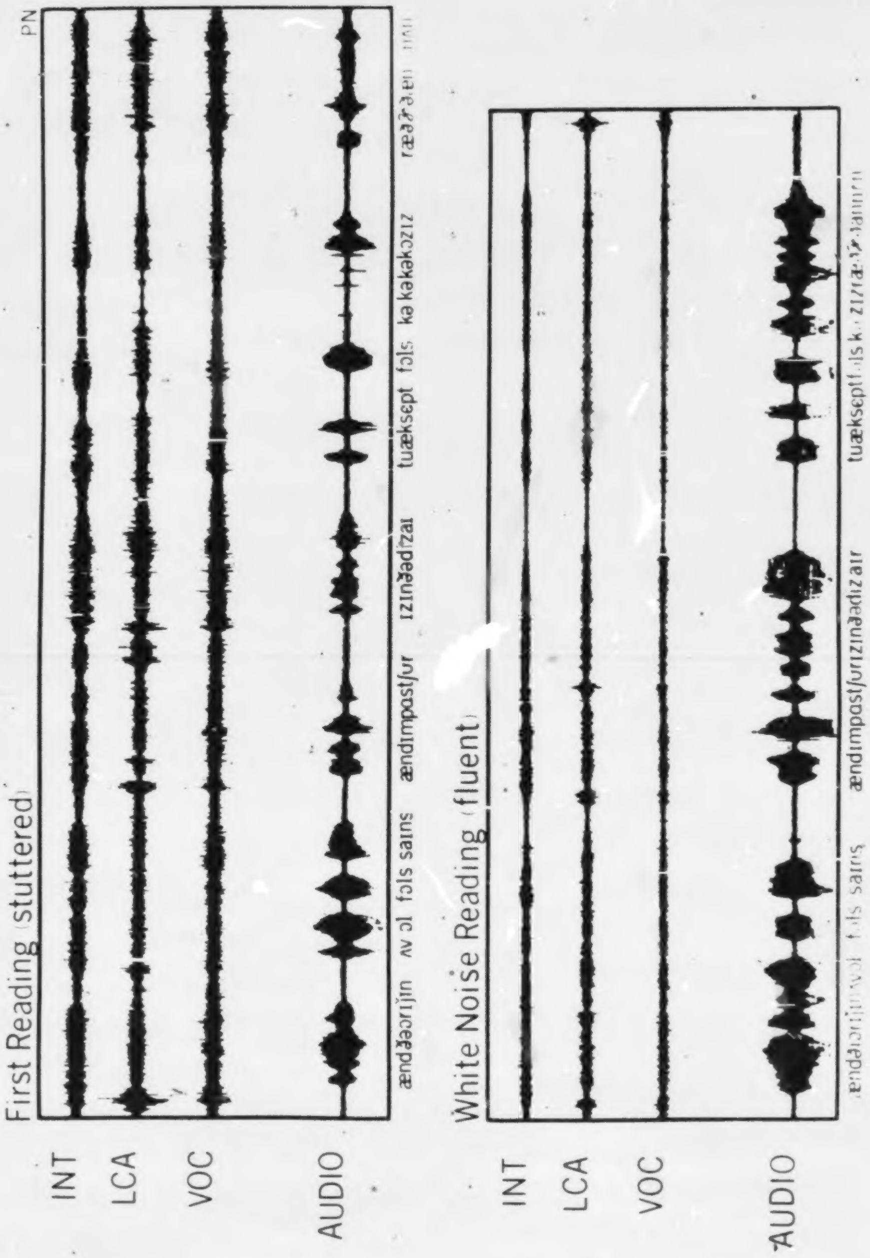

First Reading (stuttered)
INT
LCA
VOC
AUDIO
PN
White Noise Reading (fluent)
INT
LCA
VOC
AUDIO

FIGURE 2

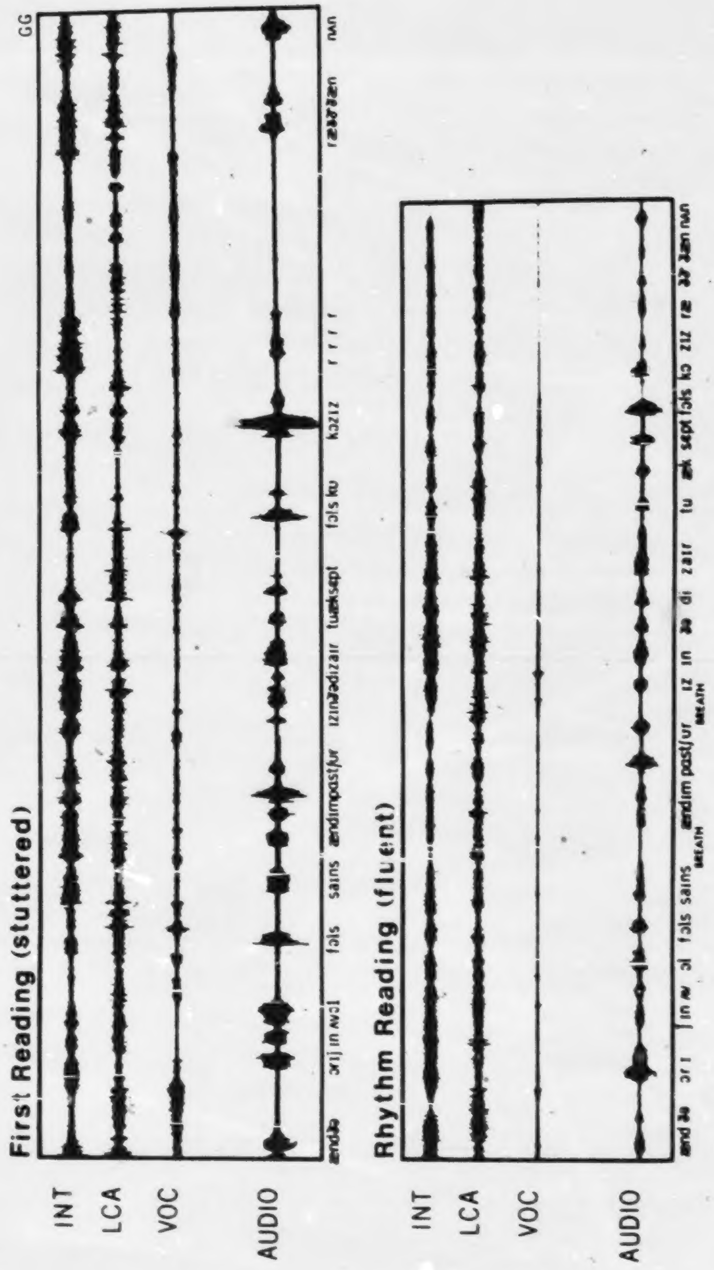

First Reading (stuttered)
INT
LCA
VOC
AUDIO
GG
Rhythm Reading (fluent)
INT
LCA
VOC
AUDIO
BREATH
BREATH

FIGURE 3

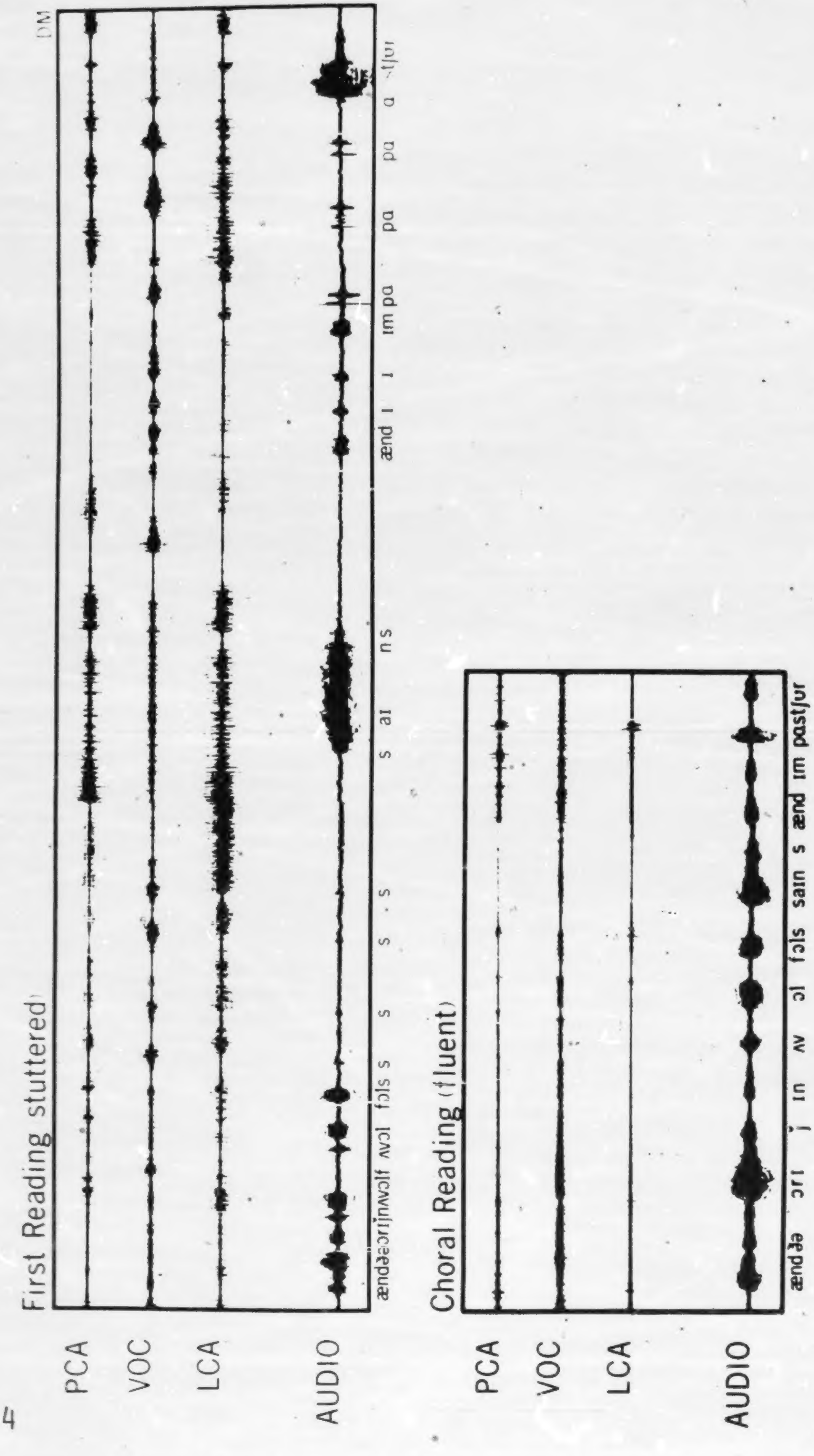

First Reading (stuttered)
PCA
VOC
LCA
AUDIO
ændəɔrɪnvəlfɪcəv fɔls s s s s s aɪ ns ænd ɪ ɪ ɪmpa pa pa pa a stʃʊr
DM
Choral Reading (fluent)
PCA
VOC
LCA
AUDIO
ændʒə ɔrɪ j ɪn ʌv ɔl fɔls saɪn s ænd ɪm pastʃʊr

FIGURE 4

These data appear to support the Wingate hypothesis. The three subjects do indeed use laryngeal activity patterns that differ from their stuttering modes when they speak under these three fluency-evoking conditions.

Further processing of the data allows us to examine the details of laryngeal articulation of individual words. However, interpretation of data on the stuttering subjects is based on analysis of laryngeal articulatory activity in normals. Therefore, a brief summary of these studies is necessary. Current research indicates that the posterior cricoarytenoid is responsible for abduction of the vocal folds with increasing levels of PCA activity correlating with width of glottal opening (Hirose and Gay, 1972; Hirose and Ushijima, 1974). Segmentally it is active for voicelessness and aspiration. The interarytenoids, lateral cricoarytenoid, and thyroarytenoid, while generally grouped together as vocal-fold adductors, exhibit activity patterns indicative of functional differentiation (Hirose and Gay, 1972; Hirose, 1974). The interarytenoids are active for all voiced sounds, vocalic and consonantal, with sharp drops in activity for voicelessness. The thyroarytenoid (vocalis) and the lateral cricoarytenoid show increasing activity for vowel segments with decreases in activity for consonant segments. The lateral cricoarytenoid applies medial compression and is very active for tight glottal closure, as in glottal stop or swallow (Hirose and Gay, 1973).[2] The thyroarytenoid increases anterior-posterior vocal-fold tension and interacts with the cricothyroid in control of fundamental frequency (Shipp and McGlone, 1971; Gay et al., 1972).

Within this framework stuttered and fluent utterances may be compared.

Figure 5 shows data on subject DM. He repeated the word "ancient" with progressive adaptation from a strong prolongation to a mild block and finally to a fluent utterance. In the first stuttered utterance, the period of prolongation is characterized by activity of the glottal abductor, the PCA, with the two adductors, the VOC and the LCA. Disruption of the normal reciprocity between abductors and adductors appears to be a critical factor. Unfortunately, this subject is the only one on whom a successful PCA recording was secured. The stuttered utterances are also marked by higher levels of activity in the adductors.

For subject PN, Figure 6 shows the word "causes," first a stuttered utterance and then a fluent utterance spoken under white noise. The first three channels are the laryngeal adductors and the fourth is the genioglossus. The peaks in the genioglossus tracing represent activity for raising the dorsum of the tongue for the /k/. Activity levels in the adductors are greater for the stuttered contrasted with the fluent utterance. It is interesting that the activity levels for the genioglossus do not show such large differences.

Figure 7 shows the same word produced nonfluently and then fluently by subject GG. The fluent utterance is spoken in rhythm. The subject repeated the

---

[2] The data showing suppression of thyroarytenoid and lateral cricoarytenoid for voiced consonants were obtained mainly on English and Japanese samples. Recent recordings of Danish and Dutch speakers show some cases of VOC activity for voiced consonants. There may be some individual or language differences that require further investigation.

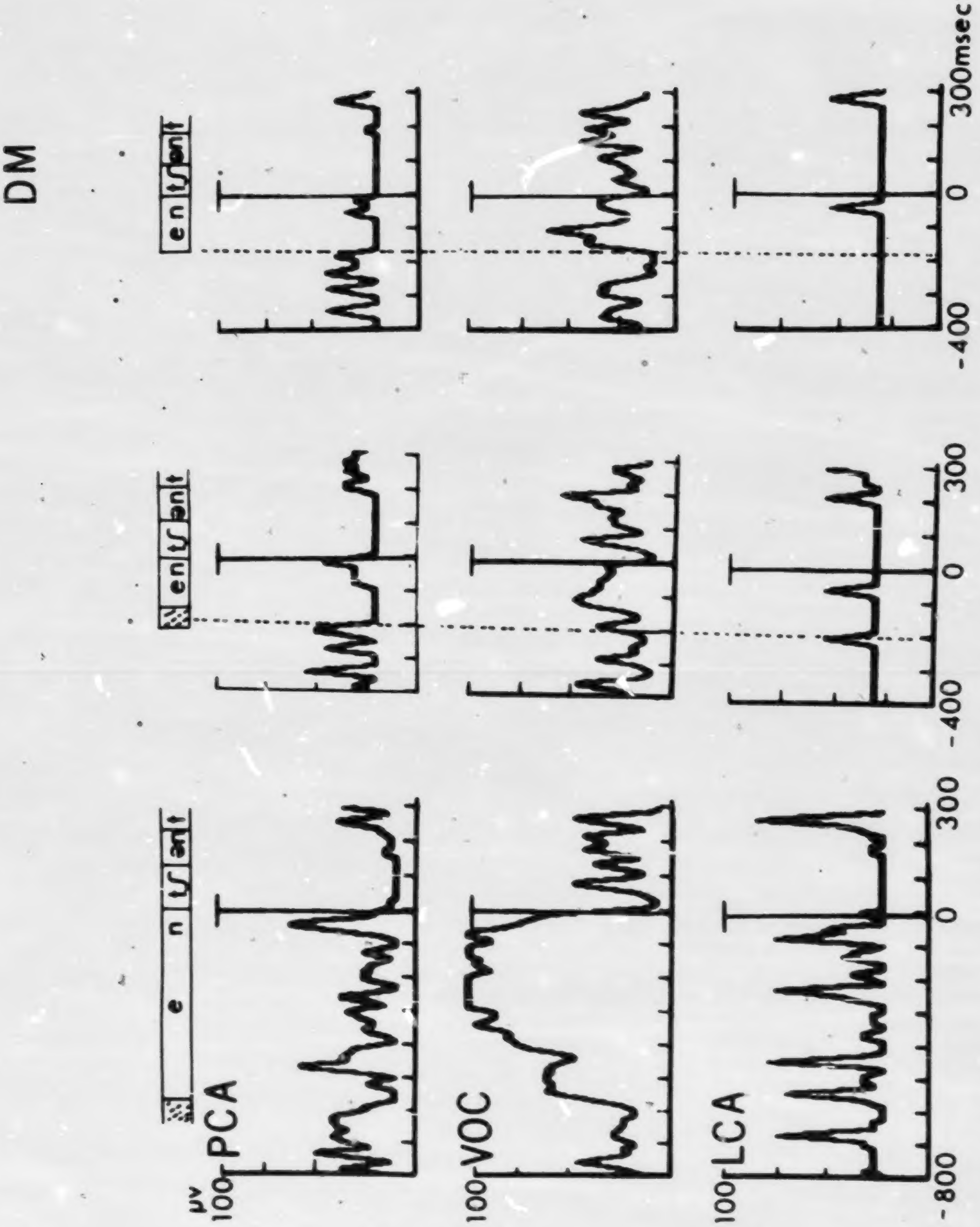

DM
μv
100
PCA
100
VOC
100
LCA
e
n
tʃ
ənt
en
tʃ
ənt
en
tʃ
ənt
-800
0
300
-400
0
300
-400
0
300msec

FIGURE 5

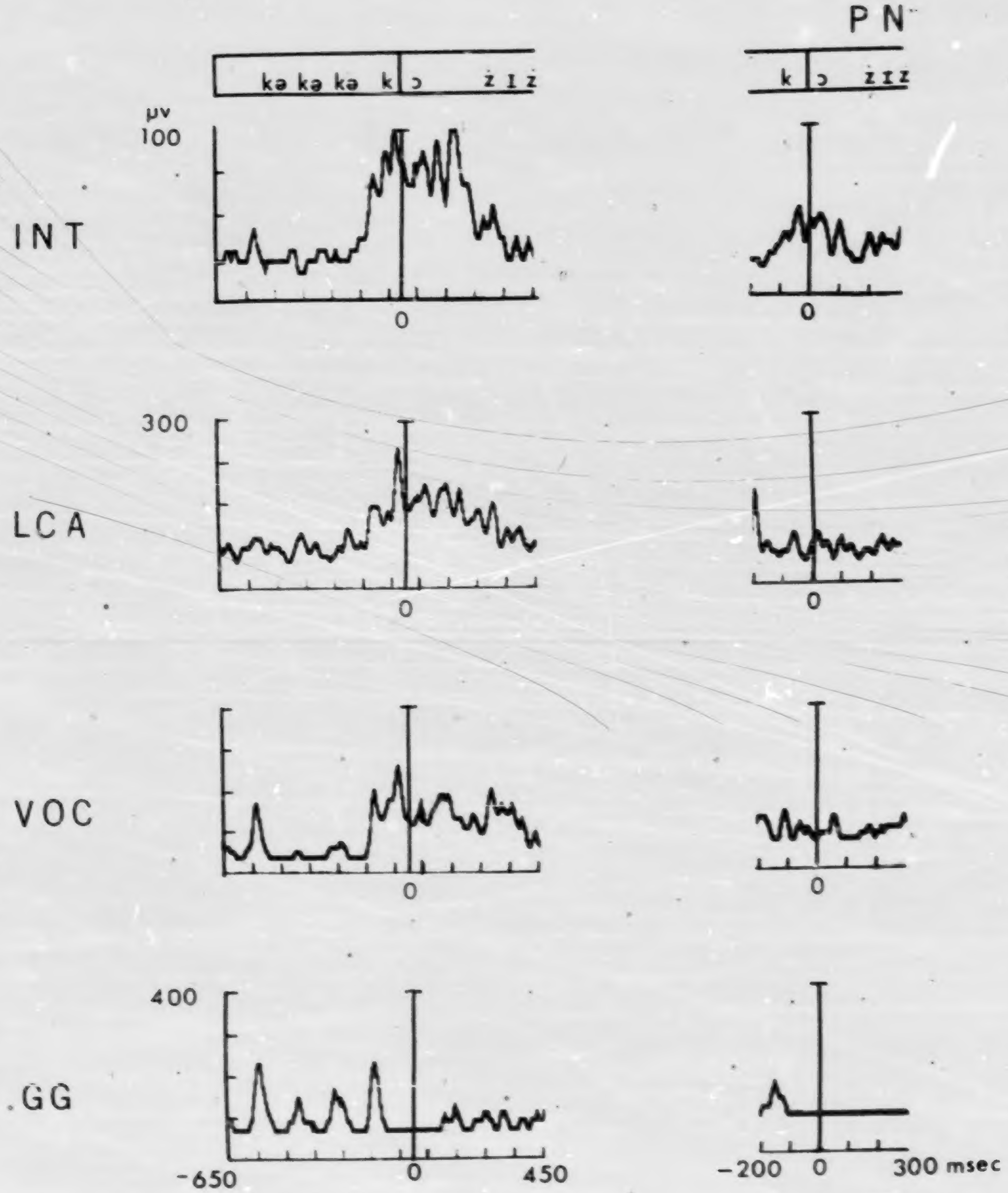

PN
kə kə kə k ɔ ż ɪ z
k ɔ ż ɪ z
µv
100
INT
0
0
300
LCA
0
0
VOC
0
0
400
GG
-650
0
450
-200
0
300 msec

FIGURE 6

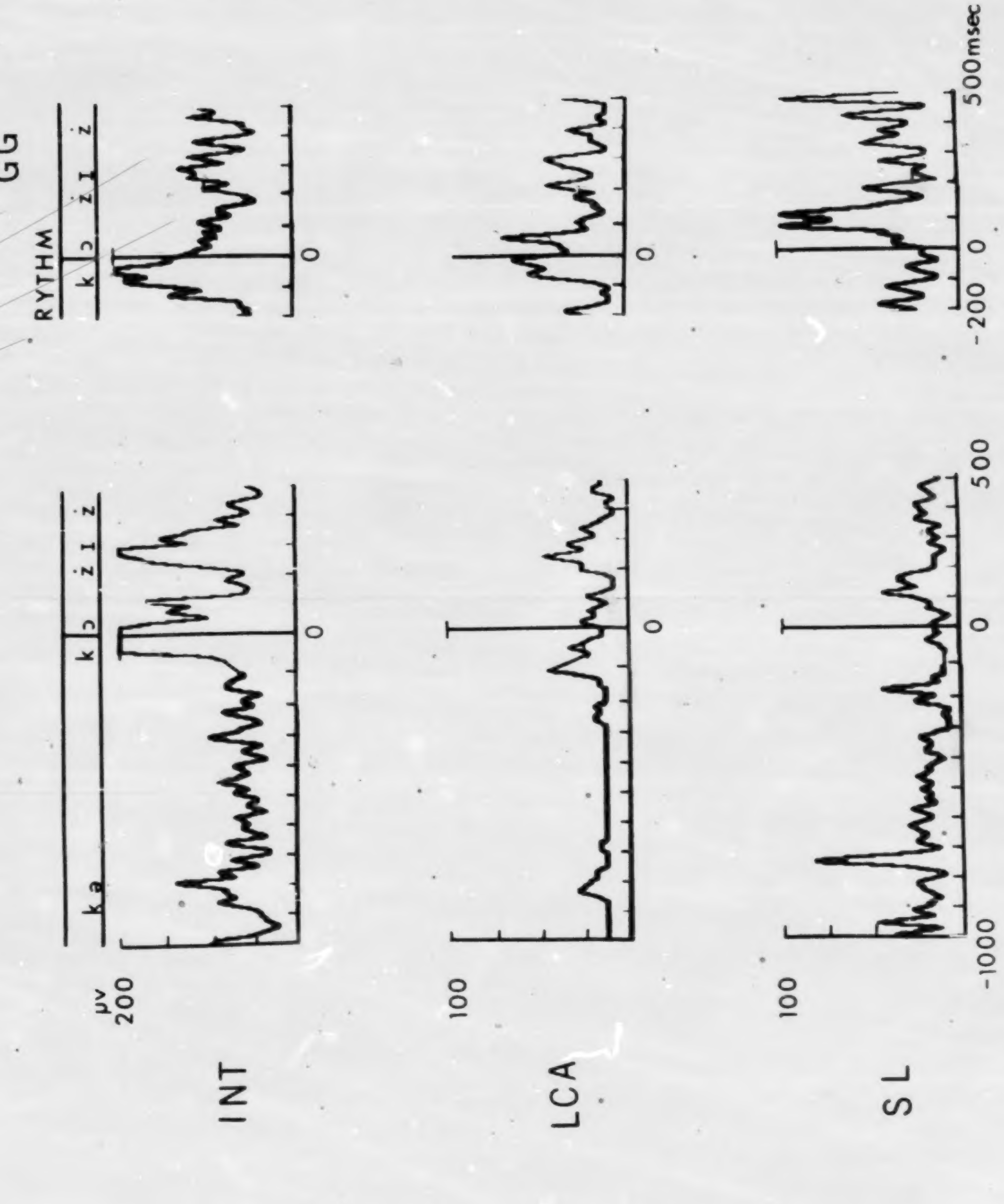

GG
RYTHM
k a
k ɔ ż ɪ ż
μv
200
100
100
INT
LCA
SL
0
-1000
500
-200
500msec

FIGURE 7

initial sound only once, paused, then uttered the word. This is typical of his pattern, which contains many "silent" blocks. The fluent utterance of the first syllable shows a synchrony of adductive activity between the interarytenoid (IA) and the LCA. In the stuttered utterance, the LCA does not act in synchrony with the IA for the vowel production. The contrast in activity in this muscle between the fluent and stuttered utterance is readily apparent. The lower graph traces the inferior longitudinal (IL), an intrinsic tongue muscle, active here for raising the tongue tip for the devoiced [z]. Note that although the [z] is not uttered in the first abortive attempt, the tongue is obviously moving into position during the stuttered utterance. This evidence of articulatory coarticulation in a stuttering block contradicts Van Riper's (1971) hypothesis concerning the absence of coarticulation in moments of stuttering, but supports the work of Hutchinson and Watkin (1974).

In conclusion, we find that moments of stuttering are characterized by patterns of laryngeal muscle activity that are not characteristic of fluent utterance and that indeed may be incompatible with normal fluent utterance. These abnormal patterns include (1) disruption of the normal reciprocity between abductors and adductors, (2) disruption of the normal synchrony between adductors, and (3) generally higher levels of activity in four of the intrinsic laryngeal muscles.

REFERENCES

Arnott, G. Niel. (1828) Elements of Physics, as reported in Hunt (1861).

Chevrie-Muller, C. (1963) A study of laryngeal function in stutterers by the glottal-graphic method. In Proc. VII Congress de la Societé Française de Medicine de la Voix et de la Parole, Paris.

Contour, E. G., D. W. Brewer, and G. N. McCall. (1974) Laryngeal activity during the moment of stuttering: Some preliminary observations. Paper presented at the annual convention of the American Speech and Hearing Association, November 7, Las Vegas, Nev.

Faaborg-Anderson, K. (1957) Electromyographic investigation of intrinsic laryngeal muscles in humans. Acta Physiol. Scand., Suppl. 41, 140.

Fujita, K. (1966) Pathophysiology of the larynx from the viewpoint of phonation. J. Japan. Soc. Otorhinolaryngol. 69, 459.

Gay, T., M. Strome, H. Hirose, and M. Sawashima. (1972) Electromyography of the intrinsic laryngeal muscles during phonation. Ann. Otol. Rhinol. Laryngol. 81, 401-408.

Hirano, M. and J. Ohala. (1969) Use of hooked-wire electrodes for electromyography of the intrinsic laryngeal muscles. J. Speech Hearing Res. 12, 362-373.

Hirano, M., J. Ohala, and W. Vennard. (1970) Regulation of register, pitch and intensity of voice. Folia Phoniat. 22, 1-20.

Hirose, H. (1971) Electromyography of the articulatory muscles: Current instrumentation and technique. Haskins Laboratories Status Report on Speech Research SR-25/26, 73-86.

Hirose, H. (1974) Functional differentiation of the glottal adductors. Japan. J. Otol. 77, 46-57.

Hirose, H. and T. Gay. (1972) The activity of the intrinsic laryngeal muscles in voicing control: An electromyographic study. Phonetica 25, 140-164.

Hirose, H. and T. Gay. (1973) Laryngeal control in vocal attack: An electromyographic study. Folia Phoniat. 25, 203-213.

Hirose, H. and T. Ushijima. (1974) The function of the posterior cricoarytenoid in speech articulation. Haskins Laboratories Status Report on Speech Research SR-37/38, 99-107.

Hunt, J. (1861) *Stammering and Stuttering: Their Nature and Treatment*, reprinted in 1967. (London: Hafner Publishing Co.).

Hutchinson, J. M. and K. L. Watkin. (1974) A preliminary investigation of lip and jaw coarticulation in stutterers. Paper presented at the annual convention of the American Speech and Hearing Association, November 7, Las Vegas, Nev.

Kamiyama, G., H. Hirose, T. Ushijima, and S. Niimi. (1965) Articulatory movements of the larynx during stuttering (a film produced at the Research Institute of Logopedics and Phoniatrics, Faculty of Medicine, University of Tokyo).

Kenyon, E. L. (1943) The etiology of stammering: The psychophysiologic facts which concern the production of speech sounds and of stammering. J. Speech Dis. 8, 337-348.

Müller, J. P. (1857) *Elements of Physiology*, trans. by Baly and reported in Hunt (1861).

Port, D. K. (1971) The EMG data system. Haskins Laboratories Status Report on Speech Research SR-25/26, 67-72.

Port, D. K. (1973) Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research SR-33, 173-184.

Port, D. K. (1974) An experimental evaluation of the EMG data processing system: Time constant choice for digital integration. Haskins Laboratories Status Report on Speech Research SR-37/38, 65-72.

Shipp, T. and R. McGlone. (1971) Laryngeal dynamics associated with voice frequency change. J. Seeech Hearing Res. 4, 761-768.

Van Riper, C. (1971) *The Nature of Stuttering*. (Englewood Cliffs, N. J.: Prentice-Hall).

Wingate, M. E. (1969) Sound and pattern in "artificial" fluency. J. Speech Hearing Res. 12, 677-686.

Wingate, M. E. (1970) Effect on stuttering of changes in audition. J. Speech Hearing Res. 13, 861-873.

II. PUBLICATIONS AND REPORTS

III. APPENDIX

PUBLICATIONS AND REPORTS

Publications and Manuscripts

Arthur S. Abramson. (1974) Experimental Phonetics in Phonology: Vowel Duration in Thai. PASAA 4, 71-90.

*James E. Cutting. (in press) The Magical Number Two and the Natural Categories of Speech and Music. In Tutorial Essays in Psychology, Vol. 1 (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

_____ (1975) Aspects of Phonological Fusion. Journal of Experimental Psychology: Human Perception and Performance 104(2), 105-120.

_____ (1975) Orienting Tasks Affect Recall Performance More Than Subjective Impressions of Ability to Recall. Psychological Reports 36, 155-158.

_____ (1974) Different Speech-Processing Mechanisms Can Be Reflected in the Results of Discrimination and Dichotic Listening Tasks. Brain and Language 1, 363-373.

_____ (1974) Two Left-Hemisphere Mechanisms in Speech Perception. Perception and Psychophysics 16(3), 601-612.

James E. Cutting and Ruth S. Day. (in press) The Perception of Stop-Liquid Clusters in Phonological Fusion. Journal of Phonetics 3.

James E. Cutting and Peter D. Eimas. (1975) Phonetic Feature Analyzers and the Processing of Speech in Infants. In The Role of Speech in Language, ed. by J. F. Kavanagh and J. E. Cutting (Cambridge, Mass.: MIT Press).

+James E. Cutting and James F. Kavanagh. (in press) On the Relationship of Speech to Language. Journal of the American Speech and Hearing Association.

James E. Cutting and Burton S. Rosner. (1974) Categories and Boundaries in Speech and Music. Perception and Psychophysics 16(3), 564-570.

C. J. Darwin. (in press) The Perception of Speech. In Handbook of Perception, Vol. 7, ed. by E. C. Carterette and M. P. Friedman (New York: Academic Press).

Michael F. Dorman, James E. Cutting, and Lawrence J. Raphael. (1975) Perception of Temporal Order in Vowel Sequences with and without Formant Transitions. Journal of Experimental Psychology: Human Perception and Performance 104(2), 121-129.

---

*To appear in SR-42/43.

+Appears in this report, SR-41.

K. S. Harris. (in press) Review of *Speech and Cortical Functioning*. *Journal of the Acoustical Society of America*.

Michael Kubovy, James E. Cutting, and Roderick McI. McGuire. (1974) Hearing with the Third Ear: Dichotic Perception of a Melody without Monaural Familiarity Cues. *Science* *186*, 272-274.

Paul Mermelstein. (1975) A Phonetic-Context Controlled Strategy for Segmentation and Phonetic Labeling of Speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing* *ASSP-23*, 79-82.

Charles C. Wood and Ruth S. Day. (1975) Failure of Selective Attention to Phonetic Segments in Consonant-Vowel Syllables. *Perception and Psychophysics* *17*(4), 346-350.

Reports and Oral Presentations

Arthur S. Abramson. Pitch and Consonant Voicing in Thai: The Phonetic Plausibility of an Historical Hypothesis. Invited talk presented at the New York Academy of Sciences, 14 April 1975.

Thomas Baer. Modeling Implications of Measurements on Excised Larynxes. Presented at the 89th meeting of the Acoustical Society of America, Austin, Texas, 11 April 1975.

F. Bell-Berti and K. S. Harris. Coarticulation in VCV and CVC Utterances: Some EMG Data. Presented at the 89th meeting of the Acoustical Society of America, Austin, Texas, 11 April 1975.

C. J. Darwin. Acoustic Memory and Speech Perception. Colloquia given at Departments of Psychology, Brown University, Providence, R. I., and Massachusetts Institute of Technology, Cambridge.

C. J. Darwin and S. A. Brady. Voicing and Juncture in Stop-Liquid Clusters. Presented at the 89th meeting of the Acoustical Society of America, Austin, Texas, 9 April 1975.

M. F. Dorman, Lawrence Raphael, A. M. Liberman, and Bruno Repp. Masking-like Phenomena in Speech Perception. Presented at the 89th meeting of the Acoustical Society of America, Austin, Texas, 10 April 1975.

Elaine Fenton, Katherine Harris, and Richard Sweet. Description of Patients with Parkinson's Disease: Preliminary Report. Presented at the annual meeting of the New York State Speech and Hearing Association, 29 April 1975.

K. S. Harris. Speaker/participant in "Interdisciplinary Perspectives of Language and Language Pathology." Annual meeting of the New York State Speech and Hearing Association, 29 April 1975.

_____ Speaker/participant in "Pitfalls of Clinical Research." Annual meeting of the New York State Speech and Hearing Association, 29 April 1975.

_____ Laryngeal Dynamics: Research with Future Clinical Applications. Presented to the American Academy of Private Practice in Speech Pathology and Audiology, Tarrytown, N. Y., 15 May 1975.

_____ Supra-Laryngeal Aspects of Voice Production. Presented at the 4th Annual Symposium: Professional Care of the Voice, Juilliard School, New York City, 25 June 1975.

Leigh Lisker, Alvin M. Liberman, David Dechowitz, and Donna M. Erickson. On Pushing the Voice-Onset-Time (VOT) About. Presented at the 89th meeting of the Acoustical Society of America, Austin, Texas, 10 April 1975.

*Lawrence J. Raphael, M. F. Dorman, and A. M. Liberman. Vowel Information Conveyed by Consonant Transitions. Presented at the 89th meeting of the Acoustical Society of America, Austin, Texas, 10 April 1975.

Bruno H. Repp. Categorical Perception, Auditory Memory, and Dichotic Interference: A "Same-Different" Reaction Time Study. Presented at the 89th meeting of the Acoustical Society of America, Austin, Texas, 10 April 1975.

D. Shankweiler. On Accounting for the Poor Recognition of Isolated Vowels. Presented at the 89th meeting of the Acoustical Society of America, Austin, Texas, 10 April 1975.

Michael Studdert-Kennedy. Auditory Sequence and Phonetic Classification. Invited paper presented at the 89th meeting of the Acoustical Society of America, Austin, Texas, 8 April 1975.

M. T. Turvey. Some Thoughts on Action with Reference to Vision. Colloquium presented at University of Minnesota, February 1975.

_____ Some Thoughts on Action with Reference to Vision. Colloquium presented at Brandeis University, Waltham, Mass., February 1975.

_____ Varieties of Memory. Colloquium presented to the Northeast Ontario Mental Health Center, Canada, March 1975.

_____ Central Processes in Visual Information Processing. Colloquium presented at the University of Toronto, Canada, March 1975.

_____ Some Thoughts on Action with Reference to Vision. Colloquium presented at the University of Toronto, Canada, March 1975.

_____ Perception as Constructed, Perception as Direct. Invited talk presented at University of Connecticut, Internal-Faculty Consortium for the Study and Teaching of Human Values, April 1975.

_____ Some Thoughts on Action with Reference to Vision. Colloquium presented at the University of Illinois, April 1975.

_____ Perspectives in Vision. Invited address at annual meeting of the Orton Society of New England, April 1975.

_____ Perspectives in Vision. Invited address at annual meeting of the Orton Society of New York, April 1975.

_____ Some Thoughts on Action with Reference to Vision. Colloquium presented at Massachusetts Institute of Technology, Cambridge, May 1975.

R. R. Verbrugge and A. M. Liberman. Context-Conditioned Adaptation of Liquids and Their Third Formant Components. Presented at the 89th meeting of the Acoustical Society of America, Austin, Texas, 10 April 1975.

APPENDIX

DDC (Defense Documentation Center) and ERIC (Educational Resources Information Center) numbers:

SR-21/22 to SR-39/40

Status Report		DDC	ERIC
SR-21/22	January - June 1970	AD 719382	ED-044-679
SR-23	July - September 1970	AD 723586	ED-052-654
SR-24	October - December 1970	AD 727616	ED-052-653
SR-25/26	January - June 1971	AD 730013	ED-056-560
SR-27	July - September 1971	AD 749339	ED-071-533
SR-28	October- December 1971	AD 742140	ED-061-837
SR-29/30	January - June 1972	AD 750001	ED-071-484
SR-31/32	July - December 1972	AD 757954	ED-077-285
SR-33	January - March 1973	AD 762373	ED-081-263
SR-34	April - June 1973	AD 766178	ED-081-295
SR-35/36	July - December 1973	AD 774799	ED-094-444
SR-37/38	January - June 1974	AD 783548	ED-094-445
SR-39/40	July - December 1974	AD A007342	ED-102-633

AD numbers may be ordered from:
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22151

ED numbers may be ordered from:
ERIC Document Reproduction Service
Leasco Information Products, Inc.
P. O. Drawer O
Bethesda, Maryland 20014

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Haskins Laboratories, Inc. 270 Crown Street New Haven, Connecticut 06510	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP N/A

3. REPORT TITLE
Haskins Laboratories Status Report on Speech Research, No. 41, January-March 1975

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Interim Scientific Report

5. AUTHOR(S) (First name, middle initial, last name)
Staff of Haskins Laboratories; Alvin M. Liberman, P.I.

6. REPORT DATE June 1975	7a. TOTAL NO. OF PAGES 235	7b. NO. OF REFS 291
8. CONTRACT OR GRANT NO. NIDR: Grant DE-01774 NICHD: Grant HD-01994 VA/PSAS Contract V101(134)P-71 ONR Contract N00014-67-A-0129-0001 ARPA/ONR Contract N00014-67-A-0129-0002 DAAB03-75-C-0419(L433) NICHD Contract NIH-71-2420 NIH/GRS: Grant RR-5596	9a. ORIGINATOR'S REPORT NUMBER(S) SR-41 (1975) 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) None	

10. DISTRIBUTION STATEMENT
Distribution of this document is unlimited.*

11. SUPPLEMENTARY NOTES N/A	12. SPONSORING MILITARY ACTIVITY See No. 8

13. ABSTRACT

This report (1 January - 31 March 1975) is one of a regular series on the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. Manuscripts cover the following topics:

-Preliminaries to a Theory of Action with Reference to Vision
-Two Questions in Dichotic Listening
-Relationship of Speech to Language
-Rise Time in Nonlinguistic Sounds and Models of Speech Perception
-Phonetic Coding of Words in Taxonomic Classification Task
-On the Front Cavity Resonance, and Its Possible Role in Speech Perception
-Synthetic Speech Comprehension: Comparison of Listener Performances with and Preferences among Different Speech Forms
-Testing Synthesis-by-Rule with OVEBORD Program
-Stress and the Elastic Syllable: Delineating Lexical Stress Patterns in Connected Speech
-VOT or First-Formant Transition Detector?
-Pitch in Perception of Voicing States in Thai: Diachronic Implications
-Facial Muscle Activity in Production of Swedish Vowels: An EMG Study
-Combined Cinefluorographic-EMG study of the Tongue during Production of /s/: Preliminary Observations
-Velar Movement and Its Motor Command
-The Stuttering Larynx: An EMG, Fiberoptic Study of Laryngeal Activity Accompanying the the Moment of Stuttering

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

UNCLASSIFIED
Security Classification

A-31408

*This document contains no information not freely available to the general public. It is distributed primarily for library use.

UNCLASSIFIED
Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Action - Vision						
Dichotic Listening						
Speech - Language						
Nonlinguistic Sounds - Speech Perception						
Phonetic Coding - Taxonomic Classification						
Resonance - Front Cavity						
Synthetic Speech - Comprehension						
Synthesis-by-Rule						
Lexical Stress in Connected Speech						
Voice Onset Time or First-Formant Transition Detector						
Pitch Perception - Thai						
Vowel Production - EMG Study of Swedish						
Fricative Production - EMG, Cine Study of /s/						
Velar Movement						
Stuttering - EMG Fiberoptic Study of Larynx						

DD FORM 1 NOV 65 1473 (BACK)
S/N 0101-807-6821

UNCLASSIFIED
Security Classification

A-31409